

Universidade de Lisboa  
Instituto de Geografia e Ordenamento do Território



**‘Big Data’ e Redes Sociais na Análise Espacial de Padrões Turísticos  
em Contexto Urbano. O caso de Lisboa**

**Luis Vicente Encalada Abarca**

Dissertação orientada pelo Prof. Doutor Jorge Rocha  
e coorientada pelo Prof. Doutor Carlos Cardoso Ferreira

Mestrado em Sistemas de Informação Geográfica e Modelação Territorial  
Aplicados ao Ordenamento

2016



Universidade de Lisboa  
Instituto de Geografia e Ordenamento do Território



**‘Big Data’ e Redes Sociais na Análise Espacial de Padrões Turísticos  
em Contexto Urbano. O Caso de Lisboa**

**Luis Vicente Encalada Abarca**

Dissertação orientada pelo Prof. Doutor Jorge Rocha  
e coorientada pelo Prof. Doutor Carlos Cardoso Ferreira

Júri:

Presidente: Professor Doutor José Manuel Henriques Simões do Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa;

Vogais:

- Professor Doutor José António Pereira Tenedório da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa;
- Professor Doutor Fernando Jorge Pedro da Silva Pinto da Rocha do Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa.

2016





## Agradecimentos

Expresso sinceros agradecimentos aos meus orientadores, Professor Doutor Jorge Rocha e Professor Doutor Carlos Cardoso Ferreira pela ajuda prestada (observações científicas e correcção do texto) e por tantos outros aspectos relacionados com o desenvolvimento desta dissertação. Quero também estender o meu agradecimento a todos os professores do mestrado em Sistemas de Informação Geográfica e Modelação Territorial Aplicados ao Ordenamento que sempre me abordaram com boas conversas e conselhos durante o meu percurso no Instituto de Geografia e Ordenamento do Território (IGOT).

Há minha família pelo suporte desde sempre. A Elena-Andreea Bicu merece uma especial menção pelo apoio moral e pela companhia que sempre me deu no decurso do mestrado e durante a minha estadia em Lisboa. A todos os Colegas um muitíssimo obrigado pela boa convivência e incentivos.

Um reconhecido agradecimento às Instituições que contribuíram para a realização desta dissertação, nomeadamente: à *Secretaría de Educación Superior, Ciencia, Tecnología e Innovación* (SENESCYT) do Equador, que outorgou o apoio económico para efectuar os estudos em Lisboa, e ao IGOT por facilitar o ingresso ao laboratório de geoprocessamento (GEOMODLAB).

*Gracias totales...*



## Resumo

O entendimento de como os turistas usam o espaço nos ambientes urbanos tem aplicações práticas na administração e planeamento do turismo nos locais de destino. Os métodos tradicionais que permitem registrar as atividades turísticas nos centros urbanos dificilmente podem produzir dados quantitativos referentes ao comportamento espacial dos turistas. Recentemente, outras fontes não oficiais, enquadradas no campo das *social media*, têm vindo a ser aproveitadas, sobretudo pelos seus dados serem de livre acesso e usufruírem de atributos de geolocalização, com elevada resolução espacial e temporal.

Os serviços de partilha de fotos, de mensagens instantâneas, de avaliação de consumidores e outras plataformas “Web 2.0”, enquadradas num ambiente colaborativo e de partilha de conteúdos, formam parte destas novas fontes que se caracterizam pelos altos níveis de participação e pela quantidade de utilizadores. Relativamente aos estudos baseados na informação extraída dos serviços de partilha de fotos, alguns autores têm referido que este tipo de dados proporciona novas oportunidades para as ciências geográficas e para o estudo de fenómenos sociais, podendo ser utilizado como *proxy* para analisar a distribuição espacial de turistas e como medida de atractividade do espaço.

Na presente dissertação, utilizaram-se as fotos (*geotagged*) partilhadas pelos utilizadores da rede ‘Panoramio’, tiradas durante a sua visita a Lisboa entre 2007 e 2014, como fonte de dados para analisar o comportamento espaciotemporal da procura turística na cidade. O conjunto total de dados (> 75.000 fotografias) foi segmentado em turistas e locais, tendo como referencia a duração do período em que cada utilizador tirou as fotografias (*i.e.*, estadia inferior a 3 dias), resultando numa amostra de 17.604 fotos.

A análise da localização das fotografias permitiu identificar os padrões espaciais dos turistas e, através do número de fotos tiradas, quantificar a capacidade de atracção dos principais pontos de interesse. A comparação entre os *clusters* espaciais das fotos dos turistas e os mapas dos principais pontos de atracção denota forte correlação, uma vez que os *clusters* coincidiram com áreas conhecidas por atrair mais turistas. Finalmente, através da análise de regressão linear múltipla, verificou-se que são os ‘monumentos’, os elementos que apresentam maior atractividade, e apurou-se que existem áreas do centro histórico, cujo potencial turístico está subestimado.

**Palavras-chave:** *Geotagged* fotos, Turismo Urbano, redes sociais, Lisboa, Análise Espacial



## Abstract

Understanding how tourists use space in urban environments has practical applications in destination management and planning. It reveals key issues for organizing facilities and essential services, creating new products, impact management, and other purposes related to urban tourism.

Alongside the development of new information and communication technologies, less traditional data sources have been considered in the analysis of urban tourism, particularly for assessing the spatial behavior of tourists. Photo-sharing and instant messaging services, consumer review websites and other Web 2.0 platforms are examples of these new sources, characterized by the prominent participation of users, high volume of user-generated content and its open access nature.

Several studies have been led using data extracted from social networks platforms, as these services provide open and well-defined interfaces to interact with. They allow clients to extract the increasingly user-generated content, often with high spatio-temporal resolution, and enriched with a set of other attributes such as geolocation, timestamps, user identification, individual profiles, etc.

Such crowdsourced data provides new insights and could be used to complement authoritative data. Considering the constraints about collecting data for urban tourism studies, some authors concluded that this information can be used as *proxy* to measure the attractiveness of places as well as the spatial distribution of tourists in the city.

In this thesis, we analyzed the spatial distribution of tourists in Lisbon based on data from ‘Panoramio’ social network between 2007 and 2014. Firstly, the data was retrieved through its respective API, filtered and merged into a relational database. Then, the data was mapped and time-space statistics were applied to find the tourists *hot spots*. Furthermore, a regression analysis method was carried out, in order to find the spatial relations between the observed pattern (geographical agglomeration of tourist’s photos) and a set of 24 selected factors (explanatory variables).

The original dataset (> 75,000 geotagged photos) was segmented into tourists and locals, using the time period between the first and last photo uploaded by each user (less than 3 nights stay), resulting in a sample of 17,604 photos taken by tourists. The analysis of photo’s locations enables us to detect the spatial patterns of tourists and to quantify, through the amount of pictures taken, the attractiveness of the major points of interest.

The comparison between the spatial clusters of tourist's photos and the map of the main tourist attractions indicated that the results are highly correlated, since the locations of clusters matched with the well-known more attractive tourist's areas. Finally, by multiple linear regression analysis, it was found that the historical elements are those that appear as most attractive spots, and revealed that there are areas inside the historic center, whose tourism potential is underestimated.

We present the conclusions for the city of Lisbon, discussing pros and cons of this emerging data source, and also some highlights and challenges of big data/social networks on urban tourism management and planning.

**Keywords:** Geotagged photos; Urban Tourism; social networks; Lisbon; spatial analysis.

# Índice geral

Agradecimentos .....	i
Resumo .....	iii
Abstract .....	v
Índice geral.....	vii
Índice de figuras .....	ix
Índice de tabelas .....	xi
<b>1. Introdução .....</b>	<b>1</b>
1.1. Objectivos .....	3
1.2. Estrutura da tese.....	4
<b>2. O Turismo Urbano: “Inovação nos processos de investigação” .....</b>	<b>5</b>
2.1. A análise do Turismo Urbano: as suas abordagens .....	6
2.2. Métodos tradicionais para a monitorização das actividades turísticas .....	7
2.3. <i>User-generated content</i> , “Big data” e redes sociais como fontes não-tradicionais de dados .....	9
2.3.1. <i>Geotagged</i> fotos de redes sociais para a análise do comportamento espaço-temporal dos turistas, no contexto urbano .....	11
2.3.2. A escolha da rede social: Panoramio .....	13
2.3.2.1. Limitações dos dados .....	15
<b>3. Análise espacial .....</b>	<b>19</b>
3.1. A conjugação da análise espacial e temporal .....	19
3.2. Conceitos fundamentais para análise espacial .....	20
3.2.1. Autocorrelação espacial .....	20
3.2.2. Vizinhança, matriz de proximidade espacial .....	21
3.2.3. A agregação dos dados.....	22
3.2.4. Métodos quantitativos de análise espacial .....	23
3.3. Análise estatística de dados espaciais .....	24
3.3.1. Exploração de dados espaciais .....	24
3.3.2. Métodos estatísticos .....	26
3.3.2.1. Características espaciais da distribuição .....	28
3.3.2.2. Índice do vizinho mais próximo .....	30
3.3.2.3. <i>K-function</i> .....	32
3.3.2.4. Análise de <i>Clusters</i> e <i>Outliers</i> (ACO).....	35
3.3.2.5. Análise de pontos quentes emergentes ( <i>Emerging hot spot analysis</i> ). .....	37
3.3.2.6. Regressão linear múltipla .....	49
<b>4. Métodos e dados.....</b>	<b>51</b>
4.1. Enquadramento geográfico: O turismo na cidade de Lisboa.....	51
4.2. Procedimento metodológico para a análise espacial.....	55

4.2.1. Informação base para o estudo .....	59
4.2.2. Programas e aplicações empregados .....	59
4.3. Extracção dos dados.....	60
4.3.1. O processo de requerimento de informação .....	60
4.3.2. A automatização do requerimento de informação .....	64
4.3.3. Extracção dos dados para o estudo.....	69
4.3.3.1. Determinação das coordenadas das subáreas .....	70
4.3.3.2. Executando o código para a extracção dos dados .....	71
4.3.3.3. Conversão dos ficheiros de resposta (texto) para folhas de cálculo....	72
4.4. Criação da base de dados e gestão da informação .....	73
4.4.1. Determinação do conjunto de dados a utilizar no estudo.....	74
4.4.1.1. Comparação da informação extraída nos distintos testes.....	74
4.4.1.2. Diferenciação das fotos tiradas por visitantes e locais .....	77
4.4.2. Migração do conjunto de dados para uma BD espacial, em ambiente SIG	80
<b>5. Análise da distribuição espacial dos turistas em Lisboa.....</b>	<b>83</b>
5.1. Pré-processamento dos dados espaciais.....	83
5.1.1. Delimitação da área de estudo.....	83
5.1.2. Agregação das fotos por freguesias .....	83
5.1.3. Agregação das fotos por unidades hexagonais .....	84
5.2. Exploração dos dados .....	87
5.2.1. A localização das fotografias dos turistas .....	88
5.2.2. Distribuição das fotos dos turistas por microcentralidade .....	98
5.2.3. Classificação de fotos segundo a data em que foram carregadas.....	99
5.2.4. Distribuição das fotos dos turistas por freguesias .....	103
5.3. Análise estatística .....	105
5.3.1. Centro médio e elipse do desvio padrão .....	107
5.3.2. <i>Nearest neighbor index</i> .....	108
5.3.3. <i>K-function (Multi-distance spatial cluster analysis)</i> .....	109
5.3.4. Análise de <i>Clusters</i> e <i>Outliers</i> .....	110
5.3.5. Análise de pontos quentes emergentes.....	111
5.3.6. Regressão linear múltipla.....	113
<b>6. Análise dos resultados .....</b>	<b>123</b>
6.1. Centro médio e elipse do desvio padrão .....	123
6.2. Índice do vizinho mais próximo (NNI).....	124
6.3. <i>K-function</i> .....	125
6.4. Análise de <i>Clusters</i> e <i>Outliers</i> .....	126
6.5. Análise de pontos quentes emergentes .....	133
6.6. Análise do potencial territorial .....	142
<b>7. Conclusões .....</b>	<b>145</b>
<b>Bibliografia.....</b>	<b>151</b>
<b>Anexos.....</b>	<b>163</b>



## Índice de figuras

Figura 3.1: Componentes do gráfico resultante da função K do Ripley.....	34
Figura 3.2: Estrutura do cubo espaço-temporal.....	38
Figura 3.3: Conjunto de dados abrangendo o período 3 de Setembro de 2015, a 12 de Setembro de 2015. Fonte: Adaptado de ArcGis <i>Help</i> . ....	39
Figura 3.4: Alinhamento espaço temporal por data de término. ....	40
Figura 3.5: Alinhamento espaço temporal por data de início.....	41
Figura 3.6: Alinhamento espaço temporal segundo um “Tempo de referência”. ....	42
Figura 3.7: Caixas de agregação e relação espaço-tempo. ....	43
Figura 3.8: Fluxo de processamento de dados da análise de pontos quentes emergentes. Fonte: Adaptado de ArcGis <i>Help</i> . ....	46
Figura 3.9: Vizinhança de Von Newman (5x5). Fonte: Adaptado de ArcGis <i>Help</i> .....	47
Figura 3.10: Diagramas de dispersão: relação positiva (a), relação negativa (b), e duas variáveis não relacionadas (c).....	50
Figura 4.1: Centralidades da Região de Lisboa. Adaptado do PETRL (2015). ....	51
Figura 4.2: Microcentralidades turísticas na cidade de Lisboa. Fonte: Plano Estratégico para o Turismo na Região de Lisboa 2015 – 2019. ....	54
Figura 4.3: Mapa cognitivo do processo de análise espacial utilizado no estudo. ....	56
Figura 4.4: Processo para extrair informação do servidor da Panoramio. Adaptado de Kádár e Gede (2013).....	61
Figura 4.5: Visualização da resposta do servidor Panoramio num editor de texto. ....	62
Figura 4.6: Malhas regulares. ....	70
Figura 4.7: Interface do programa editor WingIDE 5.1. ....	71
Figura 4.8: Interface da aplicação “Opal Converter”, para converter os ficheiros de JSON para XLS. ....	73
Figura 4.9: Tabelas importadas para a base de dados.....	74
Figura 4.10: Modelo de tabela importada para a base de dados.....	74
Figura 4.11: Tabela base de fotos da área de estudo. ....	78
Figura 4.12: Diferencia de datas das fotos carregadas pelos visitantes.....	80
Figura 4.13: Dados para o estudo organizados na base de dados espaciais. ....	82
Figura 5.1: Polígonos utilizadas para a criação de malhas regulares. ....	84
Figura 5.2: Dimensão da célula da malha regular, agregação por hexágonos.....	85
Figura 5.3: Malha de hexágonos sobre a área de estudo. ....	86
Figura 5.4: Unidades hexagonais resultantes da agregação das fotos. ....	87
Figura 5.5: Fotos dos turistas de Lisboa. ....	89
Figura 5.6: Locais de interesse, nordeste da cidade (Parque das Nações).....	90
Figura 5.7: Locais de interesse, eixo central (Praça de Comercio – Praça dos Restauradores). ....	91
Figura 5.8: Locais de interesse, eixo central (Marquês de Pombal), e Interior. ....	92
Figura 5.9: Locais de interesse, núcleo centro-este. ....	93
Figura 5.10: Locais de interesse, núcleo centro – oeste. ....	94
Figura 5.11: Locais de interesse, eixo ribeirinho (Tramo Alcântara - Belém).....	95
Figura 5.12: Locais de interesse, núcleo sul (Belém).....	96

Figura 5.13: Distribuição das fotografias pelas microcentralidades.....	99
Figura 5.14: Número de fotos de turistas extraídas do Panoramio, 2006-2014. ....	100
Figura 5.15: Classificação de fotos dos turistas por meses. ....	101
Figura 5.16: Classificação dos utilizadores por meses. ....	102
Figura 5.17: Densidade de fotos nas freguesias, classificada pelas quebras naturais. .	104
Figura 5.18: Densidade de fotos nas freguesias, classificada pelos desvios padrões..	105
Figura 5.19: Esquema processual da aplicação dos métodos de análise espacial. ....	107
Figura 5.20: Inverso da distância euclidiana à variável Igrejas.....	114
Figura 5.21: Coeficiente de correlação de <i>Pearson</i> entre as variáveis pré-seleccionadas. .....	116
Figura 5.22: <i>Clusters</i> de variáveis e zonas quentes. ....	117
Figura 5.23: Significância estatística das correlações e dos <i>clusters</i> . ....	118
Figura 6.1: Características da distribuição, Centro médio e Elipse do desvio padrão.	123
Figura 6.2: Função L(d) para as fotos agregadas dos anos a) 2007, b) 2009 e c) 2012. .....	126
Figura 6.3: <i>Clusters</i> e <i>outliers</i> espaciais do turismo em Lisboa. ....	127
Figura 6.4: Índice local de Moran. ....	128
Figura 6.5: Tipos de <i>clusters</i> , zona de Belém.....	130
Figura 6.6: Tipos de <i>clusters</i> e <i>outliers</i> , zona centro.....	130
Figura 6.7: Tipos de <i>clusters</i> , zona norte.....	132
Figura 6.8: Tipos de pontos quentes emergentes.....	135
Figura 6.9: Visualização 3D da conformação dos pontos quentes emergentes, zona centro. ....	136
Figura 6.10: Tipos de pontos quentes emergentes, zona sul da cidade. ....	137
Figura 6.11: Tipos de pontos quentes emergentes, zona centro. ....	138
Figura 6.12: Pontos quentes emergentes, zona norte.....	139
Figura 6.13: 24 horas na vida de um turista de Lisboa.....	141
Figura 6.14: Densidade de fluxos de turistas em Lisboa. ....	142
Figura 6.15: Resíduos do modelo de Regressão Linear Múltipla. ....	143

## Índice de tabelas

Tabela 3.1: Resumo do conjunto de ferramentas utilizadas na análise dos dados. ....	27
Tabela 3.2. Classes de pontos quentes e frios emergentes. Adaptado de ArcGIS Help. ....	47
Tabela 4.1: Descrição das microcentralidades turísticas de Lisboa. ....	53
Tabela 4.2: Locais de interesse e atracções visitadas, Inquérito às actividades dos turistas (OTL, 2014). ....	55
Tabela 4.3: Dados de base para a análise espacial. ....	59
Tabela 4.4: Lista dos programas utilizados no trabalho. ....	60
Tabela 4.5: Parâmetros do pedido de informação. ....	62
Tabela 4.6: Síntese do número de fotos extraídas por teste. ....	77
Tabela 5.1: Número de fotos registadas sobre os locais de interesse - atracções turísticas (por questões de funcionalidades apenas estão representados os locais com mais de 45 fotos).....	97
Tabela 5.2: Valores da distância padrão das fotografias dos turistas e locais. ....	98
Tabela 5.3: Variáveis explicativas pré-seleccionadas ....	115
Tabela 5.4: Valores de VIF.....	119
Tabela 5.5: Variáveis referidas e respectivas frequências. ....	120
Tabela 5.6: Variáveis seleccionadas para o modelo. ....	120
Tabela 5.7: Desempenho do modelo. ....	121
Tabela 6.1: Resultados índice do vizinho mais próximo.....	125
Tabela 6.2: Tendência dos dados agregados, no cubo de dados.....	133



## 1. Introdução

A presença de turistas no território traz implicações significativas a vários níveis, sobretudo nas cidades com elevada dependência desta actividade, onde a frequência dos não residentes atinge uma grande proporção do fluxo total de deslocações (Lew e McKercher, 2006) e dos níveis de concentração dos espaços (Dumbrovská e Fialová, 2014; Edward *et al.*, 2008).

O entendimento de como os turistas usam o espaço nos ambientes urbanos tem aplicações práticas na administração e planeamento do turismo nos locais de destino (Lau e McKercher, 2006; Van der Knaap, 1999), podendo aproveitar-se como contributo de suporte para a organização de instalações e dos serviços essenciais ao funcionamento da actividade, a criação de novos produtos, a gestão de impactos nas suas diversas formas, entre outros fins (Hayllar *et al.*, 2008; Sharpley e Roberts, 2005).

Na literatura, os estudos enquadrados no âmbito do turismo urbano envolvem a análise, quer da oferta, quer da procura, quer das outras componentes do “sistema”. No entanto, seja qual for a escala de análise, a abordagem comumente praticada centra o foco na oferta (Pearce, 2001).

Segundo Ashworth e Page (2011), apesar da importância quantitativa deste tipo de turismo, é curioso a pouca atenção que tem sido dada ao entendimento de como os turistas realmente usam as cidades. Tal como referido por vários autores (Kádár, 2014; Pearce, 2001), o comportamento espacial e temporal da procura (no contexto urbano) ainda não é totalmente compreendido devido à falta de dados com este nível (local) de discriminação, justificada pelas dificuldades na sua quantificação, *i.e.*, limitações na aplicação de métodos científicos para monitorização de turistas.

O nível de detalhe da estatística oficial disponibilizada por parte das autoridades competentes (*e.g.* Instituto Nacional de Estatística [INE]) ou por outras instituições públicas ou privadas (*e.g.* Câmaras Municipais, Observatórios de Turismo), em muitos os casos, não consegue um maior aprofundamento da representação da procura turística do que a contabilização dos ingressos e saídas de visitantes, quer à escala de um destino, quer à escala da divisão política do território: concelhos, distritos, regiões.

Ao contrário, os elementos da oferta (*i.e.*, hotéis, restaurantes, atracções) encontram-se melhor representados no que respeita à sua escala de interpretação, uma vez que a sua dinâmica facilita a recolha e constante actualização da informação.

Como referido por vários autores (Page, 1995; Pearce, 1999; Shoval e Isaccson, 2007), existem diversas formas de obter dados para analisar a distribuição espacial dos turistas em áreas urbanas, como por exemplo os provenientes dos repositórios dos estabelecimentos de alojamento ou de recreio (museus, jardins, parques, etc.), ou os resultantes da aplicação de procedimentos tradicionais de recolha direta como inquéritos e contagens. Contudo, no primeiro caso, estes dados não reflectem o consumo da totalidade do espaço urbano, limitando-se a apontar o local de pernoita ou de recreio (Ashworth e Tunbridge, 2000). Além disso, os dados provenientes dos alojamentos não incluem os “visitantes do dia” e geram uma procura subestimada. Todavia, existem outros espaços que fazem parte da oferta onde não se regista o número de utilizadores e, na eventualidade de se efectuar, por vezes é difícil estabelecer uma diferença entre os turistas e os habitantes locais. No segundo caso, estes procedimentos podem incluir certas restrições no que respeita ao número de agentes envolvidos (dimensão da amostra), aos custos derivados da sua execução e ao factor de actualização.

Nos últimos anos, a par do desenvolvimento das novas tecnologias de informação e de comunicação, têm vindo a surgir novas fontes de dados, menos tradicionais, que podem ser aproveitadas para o estudo do turismo urbano, particularmente, para a análise do comportamento espacial dos turistas. Os serviços de partilha de fotos, de mensagens instantâneas, de avaliação de consumidores e outras plataformas “Web 2.0” envolvidas num ambiente colaborativo de partilha de conteúdos, formam parte integrante destas novas fontes, que se caracterizam pelos altos níveis de participação e pela quantidade de informação gerada (Diaz *et al.*, 2012). Estas plataformas expandem o cenário da quantidade/diversidade de fontes disponíveis, e deixam de lado a condição em que a informação era exclusivamente produzida pelas autoridades oficiais e através de métodos científicos clássicos. Estas fontes não oficiais, enquadradas no campo das *social media*, têm promovido os próprios indivíduos como geradores de informação.

Recentemente, a informação disponível nas redes sociais tem vindo a ser aproveitada, nomeadamente, pelos seus dados serem de livre acesso e com um alto nível de resolução espacial e temporal. Este tipo de dados (*crowdsourced data*) pode complementar a informação das estatísticas oficiais (Goodchild e Li, 2012).

Assim, como referido por Buhalis e Law (2008), e Herrera *et al.* (2015), denota-se uma tendência crescimento no número de turistas que partilham as suas experiências através da publicação das suas recomendações, opiniões, fotos ou vídeos sobre um

destino, actividade ou serviço, em sítios de redes sociais. Os turistas, quando utilizam as redes sociais, deixam grandes quantidades de rastros digitais, resultando num conjunto multidimensional de dados conhecido como “*Big Data*” (Buhalis e Amaranggana, 2014).

Vários estudos têm-se focado na análise de turismo urbano com recurso a dados provenientes de redes sociais (*e.g.* Twitter, Flickr, Panoramio), uma vez que estes serviços permitem a extracção da informação publicada pelos seus utilizadores, a qual vem acompanhada de outros atributos relevantes: geo-localização (através de coordenadas geográficas), data de publicação, identificação do utilizador, etc.

Relativamente aos estudos baseados na informação extraída dos serviços de partilha de fotos (*e.g.* Flickr e Panoramio), alguns autores têm referido que este tipo de dados originados a partir da comunidade proporcionam novas oportunidades para as ciências geográficas e para o estudo de fenómenos sociais, podendo ser utilizados como *proxy* para medir a capacidade de atracção dos locais e para analisar a distribuição espacial dos turistas num determinado local de destino.

## **1.1. Objectivos**

Considerando a temática anteriormente abordada, o desenvolvimento desta dissertação prende-se, por um lado, com a análise do comportamento espaço-temporal dos visitantes da cidade de Lisboa através da aplicação de métodos de estatística espacial e temporal, com recurso a dados extraídos da rede social Panoramio e, por outro lado, com a exploração das relações espaciais que podem explicar o padrão observado nos turistas. Pretende-se com este trabalho produzir informação que contribua para um melhor entendimento da actividade turística na cidade de Lisboa.

Para atingir esta pretensão foi necessário cumprir com os seguintes objectivos específicos:

- Desenvolver um processo automático para a extracção de informação da rede social Panoramio;
- Identificar quais os locais de concentração dos visitantes e a sua variação ao longo do período de análise;
- Determinar quais os factores que melhor poderão justificar os locais de concentração.

## 1.2. Estrutura da tese

A dissertação estrutura-se em sete capítulos. Os primeiros três capítulos são de teor teórico e destinam-se à apresentação da tese e ao enquadramento conceptual do tema a tratar. Os dois capítulos seguintes são de carácter mais operacional onde se apresenta a metodologia e os procedimentos executados em todas as etapas de trabalho. Por fim, nos últimos dois capítulos, discutem-se os resultados e retiram-se as devidas conclusões.

De uma forma mais pormenorizada, o capítulo 1, trata de apresentar e explicar a pertinência da dissertação, assim como os objectivos de trabalho e a estrutura da tese.

O capítulo 2, inicia-se com o enquadramento teórico da dissertação. Nesta parte, o enfoque é colocado nas questões relacionadas com a análise do turismo urbano, as suas diversas abordagens e à necessidade de desenvolver estudos sob a perspectiva da procura. São também apresentadas as vantagens da utilização da informação proveniente de redes sociais como informação complementar para o entendimento do fenómeno turístico.

No capítulo 3, continua-se com o enquadramento teórico mas, neste caso, focado na fundamentação conceptual dos métodos estatísticos aplicados na análise da distribuição e das relações espaciais dos visitantes.

No capítulo 4 expõe-se a metodologia desenvolvida para a consecução dos objectivos. Nesta parte, descrevem-se também o procedimento para a extracção da informação e os detalhes sobre o pré-processamento dos dados, incluindo a estratégia de diferenciação entre as fotografias tiradas por turistas e locais.

No capítulo 5 explica-se a operacionalização dos métodos estatísticos no contexto dos Sistemas de informação geográfica (SIG) e efectua-se uma análise exploratória que procura discernir características gerais dos dados. Aqui, abre-se a discussão sobre a distribuição espacial e temporal dos visitantes, e das relações espaciais entre o padrão observado e as variáveis que o poderão explicar.

No capítulo 6 são expostos os *outputs* decorrentes da aplicação dos métodos estatísticos, sendo analisados de forma a responder as questões sobre o comportamento espaço-temporal dos turistas, e sobre os factores que influenciam a formação do padrão espacial dos visitantes.

Finalmente, no capítulo 7, são tecidas as conclusões do trabalho.



## 2. O Turismo Urbano: “Inovação nos processos de investigação”

O turismo urbano, como o próprio nome sugere, caracteriza-se pela prática da actividade turística em contexto urbano, sendo a cidade e o seu ambiente, o factor diferenciador dos outros tipos de turismo. Para Edwards *et al.* (2008), o turismo é uma das muitas forças sociais e económicas no ambiente urbano, e definem-no:

*“It encompasses an industry that manages a variety of products and experiences to people who have a wide range of motivations, preferences and cultural perspectives and are involved in a dialectic engagement with the host community”* (p. 1.038).

As áreas urbanas são lugares complexos, caracterizados pela sua alta densidade de infra-estruturas, pessoas e funções; pela sua heterogeneidade social e cultural; e pela sua multifuncionalidade económica (Pearce, 2001). O turismo “consome” quantidades substanciais do espaço das cidades, sendo os núcleos históricos, os diferentes tipos de museus, os parques temáticos, as praças e os jardins, parte dos elementos que acomodam esse “consumo” (Gospodini, 2001).

A dinâmica do turismo urbano não se distribui homogeneamente por toda a cidade, concentrando-se com maior intensidade em determinadas áreas, conhecidas na literatura, como “enclaves turísticos”, “distritos turísticos” ou “recintos turísticos” (Judd, 2003; Pearce, 2001; Hayllar *et al.*, 2008). De acordo com Hayllar e Griffin (2005, p. 517) *“A tourism precinct represents a distinctive geographic area within a larger urban area, characterized by a concentration of tourist related land uses, activities and visitation, with fairly definable boundaries”*.

Os turistas visitam as cidades, quer como atracções primárias ou como fornecedoras de serviços de apoio, especialmente, de transporte e de alojamento (Ashworth e Page, 2011; Darcy e Small, 2008). Estes deslocam-se à procura dos locais de interesse, das atracções (*e.g.* monumentos, museus, praças, lojas especializadas, etc.) e das experiências, tais como os eventos culturais, tradicionais e desportivos, a gastronomia, e a interacção com outras pessoas, etc. (Sheller e Urry, 2009; Ashworth e Tunbridge, 2000; Page, 1995).

No contexto urbano, o turismo não é um factor isolado, como referido por Ashworth (2012). Os turistas usam quase todos os equipamentos e serviços urbanos, alguns mais intensivamente do que outros. Não há duas cidades distintas, a do turista e a do residente, contudo, esta segmentação é muitas vezes útil para actos de gestão.

## 2.1. A análise do Turismo Urbano: as suas abordagens

Pearce (2001) sugere que, para analisar o turismo urbano, se devem considerar dois aspectos: i) a escala (*e.g.* local, distrital, municipal, regional, nacional e internacional); e ii) a temática (*e.g.* procura, oferta, desenvolvimento e impactos), e afirma:

*“Whatever the scale of analysis, the most common approach to date has been to adopt a supply-side focus based on inventories of tourism product, usually accommodation, sometimes attractions and transport and, more rarely, a combination of various elements”* (p. 929).

Sustentando o anteriormente referido, Edwards *et al.* (2008) manifestam que, a partir da década de 1990<sup>1</sup>, foram geradas uma série de contribuições que suportaram um melhor entendimento do turismo urbano, embora, maioritariamente, concebidas sobre a perspectiva da oferta. Apesar do surgimento de inúmeros trabalhos, ainda se denota um longo caminho por percorrer no que respeita ao estudo do comportamento da procura. Neste contexto, Ashworth e Page (2011) enfatizam a pouca atenção que tem sido dada à questão sobre “como os turistas realmente usam as cidades”.

O enfoque particular dos estudos baseados no *supply-side* tem sido mais explorado, por um lado, pela facilidade em identificar os seus elementos (hotéis, restaurantes, atracções, etc.) e, por outro lado, pela complexidade em recolher dados sobre o comportamento espaço-temporal dos turistas quando visitam um destino. A análise da procura, por exemplo, tem sido limitada pela indisponibilidade de dados adequados à escala, uma vez que muitas das estatísticas oficiais são recolhidas apenas ao nível de cidade (Pearce, 2001), e os métodos convencionais para registrar actividades turísticas, na maioria dos casos, não produzem dados quantitativos (Kádár, 2014).

Uma melhor leitura do turismo urbano será viável quando a informação da oferta for complementada pelos dados da procura. O entendimento dos padrões do comportamento espaço-temporal dos turistas poderá suportar as práticas da indústria, entre outros fins, para a gestão dos atractivos, a optimização dos serviços de apoio e a criação de produtos (Lau e McKercher, 2006; Van der Knaap, 1999).

---

<sup>1</sup> A análise do turismo urbano apareceu como um tema particular nos anos 1990, recebendo maior atenção por parte dos investigadores, devido à crescente procura de turistas nas cidades, e a sua promoção como uma indústria estratégica para a revitalização urbana nas cidades pós-industriais (Pearce, 2001).

## 2.2. Métodos tradicionais para a monitorização das actividades turísticas

Andar a pé, fazer compras, visitar museus e monumentos, comer num restaurante local e outras actividades são, comumente, enquadradas no âmbito do turismo urbano. Ao nível local, os visitantes deslocam-se entre os vários pontos de interesse, ou de uma actividade para outra. Conhecer quais os locais de concentração dos turistas e o momento da sua visita (estações, duração da estadia, etc.), poderia levar a uma melhor compreensão da dinâmica turística.

A mobilidade dos turistas, no contexto urbano, pode ser monitorizada através de vários métodos, uns mais tradicionais e outros menos convencionais. Shoval e Isaacson (2010) fazem uma compilação sobre os distintos métodos para recolha de dados espaço-temporais.

Os autores classificam os métodos em dois grandes grupos: i) os de observação directa e ii) indirecta. O grupo de observação directa subdivide-se em métodos de observação participativa (*participant-observer methods*) e não participativa (*non-participatory methods*). Em termos práticos, no primeiro, o investigador (em pessoa) acompanha aos indivíduos sob escrutínio. Como alternativa, no segundo caso, o investigador segue os sujeitos a uma certa distância, registrando os padrões das actividades durante um determinado período. Acerca deste método (i), os autores referem algumas desvantagens relacionadas com os custos da sua aplicação e os limites no tempo de observação.

Relativamente aos métodos de observação indirecta (*Non-observational techniques*), o mais utilizado nas ciências sociais para a recolha de informação espaço-temporal e, em particular, nos estudos de turismo, é o método denominado *time-space budget*, que consiste no registo sistemático do uso do tempo dos turistas, descrevendo a sequência e a duração das actividades, normalmente, por um período que varia de um dia para uma semana. A informação é registrada pelos próprios sujeitos, ou levantada através de inquéritos. Neste caso, as desvantagens estão relacionadas com o tempo que implica a sua operacionalização, com os custos da sua aplicação e com a precisão da informação, uma vez que os inquiridos, nas suas respostas, tendem a não pormenorizar, com exactidão, as suas actividades. Quando são os turistas, os próprios colaboradores a documentar as actividades, o principal problema, é que tanto a qualidade

como a quantidade da informação recolhida depende da capacidade do sujeito para registrar os eventos com algum grau de precisão e detalhe.

Um outro método mais acessível (no que respeita aos custos da sua aplicação), é a observação remota, *i.e.* à distância ou indirecta (ii), usado para registrar os fluxos turísticos de forma agregada, através de fotografia ou de vídeo-filmagem. Contudo, ao se registrar às pessoas de forma agregada, não se consegue discriminar os turistas do total da população. Para além disto, a extensão da área de trabalho está limitada pelo espaço que pode ser coberto pelos dispositivos e, à semelhança do método anterior, não se obtém informação sobre as motivações subjacentes (dos turistas) ligadas às actividades que estão a ser documentadas.

Por outro lado, actualmente existem outras alternativas digitais utilizadas na recolha de dados relativos ao comportamento espacial dos turistas. Nestas, ao contrário dos métodos anteriores, a informação sobre as deslocações dos turistas é extraída a partir de dispositivos *Global Positioning System* (GPS), ou através de técnicas de rastreamento que utilizam dados captados por antenas terrestres (*e.g.* rede de telemóveis). Os dispositivos GPS são confiáveis e acessíveis, tendo sido utilizados em inúmeros estudos desde que foi aberto o sistema de navegação para uso civil. Das experiencias analisadas pelos autores, sobressaem alguns pontos positivos relacionados com a resolução espacial dos dados, a capacidade de obter informação complementar (velocidade, altitude, data - hora) e a possibilidade de estender o período de recolha. No entanto, a desvantagem recai nas restrições quanto ao total de sujeitos que podem ser envolvidos nos estudos (devido ao número de dispositivos disponíveis). Esta limitação pode ser ultrapassada no caso dos dados obtidos a partir da rede de telemóveis.

Considerando a prevalência das redes celulares e a ubiquidade dos telemóveis (frequentemente utilizados, mesmo durante os períodos de férias, domésticas ou internacionais), a técnica apresenta-se como uma das mais favoráveis quando se pretende obter grandes volumes de dados. A precisão na localização das unidades (ou melhor, dos indivíduos) depende essencialmente das características da infra-estrutura da rede. Segundo os autores, este tipo de dados são mais adequados para analisar a mobilidade a nível regional, já que nas áreas urbanas o desvio, no posicionamento (localização das unidades), pode chegar até aos 100 metros devido à obstrução e/ou efeito de ricochete que o edificado introduz no sinal.

Os métodos antes descritos, em muitos os casos, são aplicados de forma conjunta (*e.g.* monitorização das deslocações através de um dispositivo GPS, e aplicação de inquéritos para retirar informação adicional sobre os indivíduos, motivações, percepções, etc.), conseguindo assim, sobrelevar as suas limitações e melhorar a qualidade da informação.

Apesar de terem sido revistas com maior ênfase, as desvantagens dos métodos, não se pretende com isto, cataloga-los como pouco eficazes, uma vez que, dependendo das necessidades de informação, qualquer um deles pode produzir resultados satisfatórios no processo de recolha de dados.

### **2.3. User-generated content, “Big data” e redes sociais como fontes não-tradicionais de dados**

O surgimento da Web 2.0 revolucionou o uso da Internet como canal de comunicação (o termo "Web 2.0" inclui uma vasta colecção de plataformas digitais, entre elas, as identificadas como *social media*). Graham (2005) define como *social media*, uma qualquer plataforma digital onde os utilizadores podem participar, criar e partilhar conteúdo. Kaplan e Haenlein (2010) distinguem os seguintes meios: *blogs*, *content communities*, redes sociais, sites de recomendações e avaliações (*Consumer review websites*), sites de mensagens instantâneas e de partilha de fotos, etc. (*e.g.* Viajecomigo, Tripadvisor, Twitter, Facebook, Flickr, Panoramio).

A utilização generalizada da Internet aumentou a influência do conteúdo partilhado nestes meios sobre o comportamento dos utilizadores e, mais especificamente, sobre o comportamento dos turistas (MacKay e Vogt, 2012; Tussyadiah, 2012). O seu impacto tem sido significativo na indústria do turismo (Leung *et al.*, 2013), denotando-se uma tendência crescente dos turistas em partilharem as suas experiências com outras pessoas através da publicação das suas recomendações, opiniões, fotos ou vídeos sobre um destino, actividade ou serviço, particularmente, em sítios de redes sociais (Buhalis e Law, 2008; Herrera *et al.* 2015).

Com o desenvolvimento e popularização da Web 2.0, está-se a experienciar uma nova “era”, onde a informação é produzida, em parte, pelos utilizadores (Bruns, 2007). Este tipo de informação é denominada, na literatura, como *User-generated content* (UGC) ou *Crowd-sourced data* (Kaplan e Haenlein, 2010), *Volunteer geographic*

*information* (VGI) - mais utilizado no campo da geografia - ou *Community-contributed data* (Goodchild, 2007; Andrienko *et al.*, 2009).

No contexto da informação geográfica, o aparecimento destas novas plataformas vieram expandir o cenário de fontes disponíveis, e distancia-nos da condição em que a informação era produzida e distribuída, exclusivamente, por parte das autoridades oficiais - *authoritative data* - (Goodchild, 2008; Sui *et al.*, 2013). As *social media* e, em particular, as redes sociais, estão cada vez mais equipadas com ferramentas de posicionamento e de mapeamento. O conteúdo criado pelos utilizadores (ao contrário das metodologias *top-down*) tem promovido os próprios indivíduos como geradores de informação (com elevada resolução espacial e temporal), ampliando o quadro de alternativas para a sua localização (Sui e Goodchild, 2011). Os turistas quando utilizam os seus telemóveis, os seus cartões de crédito, ou através do acesso às redes sociais, deixam grandes quantidades de rastros digitais sobre as suas actividades num determinado destino, resultando num conjunto multidimensional de dados conhecido como “*Big Data*”<sup>2</sup> (Buhalis e Amaranggana, 2014; Hawelka *et al.*, 2014).

Uma característica valiosa do UGC produzido nas *social media*, é que está disponível quase em tempo real, o que significa, na maioria dos casos, a informação pode ser empregue para analisar temas actuais, que requerem uma revisão constante, permitindo mudar o sentido analítico de uma abordagem estática para um processo mais dinâmico de monitorização (Sui e Goodchild, 2011).

Do ponto de vista geográfico, esta informação é de ampla cobertura, georreferenciada e, às vezes, mais oportuna do que a encontrada nas fontes oficiais (*e.g.* dados necessários durante uma emergência). Para além disso, é gratuita (aberta à comunidade) ou, noutros casos, está sujeita a meios de produção de baixo custo (Diaz *et al.*, 2012).

Por outro lado, uma clara limitação é a falta de garantia no que respeita à sua qualidade, contrariamente ao que acontece com a informação procedente de fontes oficiais, uma vez que esta última é recolhida e documentada através de procedimentos bem estabelecidos e seguindo regras estritas. No entanto, apesar da limitação supramencionada, este tipo de informação pode desempenhar um papel útil, sobretudo

---

<sup>2</sup> Este termo é utilizado para referir o grande volume de informação digital disponível em forma estruturada e não estruturada, integrado a partir de múltiplas fontes de informação (Buhalis e Amaranggana, 2014).

quando se trata da etapa inicial do estudo de um fenómeno (*i.e.* análise exploratória, geração de hipótese) (Sui e Goodchild, 2011; Goodchild e Li, 2012).

Para Goodchild e Li (2012), estes novos dados não são um substituto dos dados oficiais ou dos dados recolhidos através de métodos científicos clássicos, mas sim complementares, podendo auxiliar o processo de tomada de decisão, em cenários de monitorização multidisciplinares.

### ***2.3.1. Geotagged fotos de redes sociais para a análise do comportamento espaço-temporal dos turistas, no contexto urbano***

Durante a sua visita, os turistas, em função das suas motivações, vão consumir as experiências oferecidas pela cidade. Ainda que não se verifique um consumo activo, haverá pelo menos um consumo indiciário, através do acto de tirar fotografias (Kádár, 2014).

Assim como referido por Urry e Larsen (2011), a maioria dos turistas tiram fotografias com o objectivo de criar registos da sua visita:

*“While the tourist is away, this then moves to tracking down and capturing those images for oneself. And it ends up with travellers demonstrating that they really have been there by showing it to friends and family (...). A photograph thus furnishes evidence that someone really was there (...)”* (p. 179).

As fotografias e as viagens estão ligadas intrinsecamente. Através da fotografia os turistas capturam a natureza efémera do momento, da sua experiência, mostrando-a ao ‘mundo’ como uma representação directa da realidade, um “verdadeiro” reflexo dos lugares, pessoas e eventos (Donaire *et al.*, 2014).

A fotografia proporciona uma oportunidade para os turistas partilharem as suas experiências com outros indivíduos. Tradicionalmente, a partilha das fotos de viagem era um esforço privado, com imagens montadas em álbuns e exibidas apenas a um público restrito. Contudo, os meios digitais de partilha de conteúdos transformaram esta prática no que ela é actualmente, em grande parte, uma exposição de domínio público. As fotografias são agora amplamente produzidas, consumidas e partilhadas via internet com recurso a computadores e telemóveis e, especialmente através das redes sociais (Lo *et al.*, 2011).

A partilha de fotos *online* é comum entre os viajantes, sendo as plataformas Facebook, Flickr, Panoramio e Picasa, as utilizadas com mais frequência. Como demonstrado por Lo *et al.* (2011), no estudo que elaboraram sobre o uso de tecnologias *online* para a partilha de fotos de viagens<sup>3</sup>, aproximadamente 89% dos viajantes em lazer tiram fotografias, e desse grupo, 41% colocaram as suas fotos *online*, sendo os sites de redes sociais, de mensagens instantâneas, e *blogs* pessoais, os meios de comunicação comumente usados.

Koerbitz *et al.* (2013), analisando algumas regiões e cidades da Áustria, demonstraram que os dados extraídos de redes sociais, neste caso do Flickr, são representativos do número real de turistas. Os dados da rede foram comparados com o número de dormidas nos respectivos destinos, e determinaram que tanto ao nível das regiões como das cidades em estudo, ambos os conjuntos de dados apresentavam uma correlação forte positiva.

Kádár (2014) também fez para 16 cidades de Europa, uma comparação dos dados extraídos da mesma rede com as estatísticas do turismo dessas cidades (número de chegadas e número de dormidas), determinando, em conformidade com o caso anterior, que existe uma forte correlação positiva entre estes conjuntos de variáveis. Uma análise semelhante, onde se comparam os dados recolhidos através da rede Panoramio, com as estatísticas provenientes do sector hoteleiro (taxas de ocupação) e com informação do inquérito às actividades dos turistas na cidade de Lisboa, é apresentada na secção 5.2.3.

No contexto urbano, vários estudos têm utilizado *geotagged*<sup>4</sup> fotos (e a informação relacionada a elas) provenientes de redes sociais para analisar o comportamento dos turistas. Mais especificamente, para perceber os padrões nos movimentos intra-destino e identificar os locais de concentração (Kádár e Gede, 2013; Koerbitz *et al.*, 2013; García-Palomares *et al.*, 2015). Para além disso, a informação relativa à data em que foram realizadas as fotos, ou quando foram carregadas (*upload*) na rede, tem permitido que a análise de padrões não seja efectuada, apenas, considerando a

---

<sup>3</sup> O estudo baseou-se em informação recolhida através de inquéritos aplicados a residentes de Hong Kong, no ano 2009.

<sup>4</sup> *Geotagging* (também conhecido como Geo-marca) é o processo de adicionar metadados de identificação geográfica a diversos meios marcados geograficamente, *e.g.* fotografias, vídeo, sites, mensagens SMS, Código QR ou feeds de RSS. Pode ser encarado como uma forma de metadados geoespaciais. Neste caso é o processo onde se atribui informação de contexto geo-espacial às fotos. Na rede social Panoramio, esta informação é inserida em forma de coordenadas geográficas ou através da sua localização sobre um mapa base.



distribuição espacial das fotografias mas também a sua componente temporal (Girardin *et al.*, 2008; Andrienko *et al.*, 2009; Kisilevich *et al.*, 2010).

Diversos autores utilizaram, por exemplo, dados extraídos da rede de partilha de fotos Panoramio, em conjunto com bases de dados *online* (wikipedia, wikitravel e Foursquare) para identificar e avaliar os pontos de interesse turístico, em diferentes áreas do mundo (Tammet *et al.*, 2013). Lu *et al.*, (2010) utilizaram a informação associada a 20 milhões de fotos do Panoramio para sugerir planos de rotas de viagem personalizados. Os dados serviram para recomendar os destinos mais populares para visitar, a ordem de visita e o tempo despendido em cada destino, o caminho turístico típico dentro de cada destino, etc.

Em termos gerais, todos os autores anteriormente referidos concluíram que os dados disponíveis nestes serviços de partilha de fotos mostram um elevado potencial para o estudo do turismo urbano (da perspectiva da procura), particularmente, para encontrar padrões na distribuição espacial e temporal dos visitantes.

### **2.3.2. A escolha da rede social: Panoramio**

Os sítios de partilha de fotos surgiram a partir do ano 2004. O seu impacto social, juntamente com o incremento da informação geográfica (através dos *geotags*) em tais aplicações, têm motivado aos investigadores a considerá-los como importantes fontes de dados geográficos (Antoniou *et al.*, 2010).

Como já mencionado, existem varias plataformas digitais de partilha de fotos (Flickr, Instagram, Panoramio, Picasa Web, Geograph). Contudo, nem todas disponibilizam livremente o seu conteúdo, ou apenas contêm informação de uma área específica (*e.g.* Geograph está disponível exclusivamente para a Inglaterra e a Irlanda). Na literatura relacionada com a análise do turismo e da mobilidade, as duas redes mais utilizadas como fontes de dados são o Flickr e o Panoramio, em grande parte, porque proporcionam um acesso livre à informação partilhada pelos utilizadores.

Ambas as duas redes de partilha de fotos permitem a extracção de informação através das suas respectivas *application programming interfaces* (API). Uma API é uma interface que permite aos “clientes” interagir com o serviço *Web*, enviar pedidos e receber informação alojada na base de dados em questão. Na secção 4.3.1, descreve-se com maior

detalhe quais os dados disponíveis na rede Panoramio e o procedimento necessário para efectuar a sua extracção.

O Flickr e o Panoramio são redes sociais que permitem publicar e partilhar fotos, adicionar descrições, e avaliar e efectuar comentários sobre as imagens publicadas. Apesar das suas funcionalidades (como plataformas Web 2.0.) serem similares, têm algumas diferenças no que diz respeito ao interesse sobre o conteúdo que procuram. Muito embora no Flickr, os utilizadores possam publicar fotos de qualquer conteúdo (e.g. eventos sociais, de cariz pessoal, lugares, objectos, etc.), no Panoramio isso já não se verifica. Este ultimo ‘pede’ aos utilizadores que publiquem fotografias que ilustrem lugares, particularmente em espaços exteriores (e.g. paisagens, monumentos, ruas, edifícios, parques, e afins):

*“Panoramio is different from other photo-sharing sites because the photos illustrate places. As you browse Panoramio, notice that there aren't many photos of friends and family posing in front of places – Panoramio is all about seeing the world.”* (Em Panoramio *help*<sup>5</sup>).

O Panormio<sup>6</sup> é um sítio mantido pela comunidade para explorar lugares através da fotografia (cidades, maravilhas naturais, etc.). Em 2007, foi adquirido pela *Google* e, desde então, o número de *uploads* tem aumentado exponencialmente, até ter excedido os 100 milhões de imagens, em Dezembro de 2013 (García-Palomares *et al.*, 2015).

Em termos de visualização, as imagens não só estão disponíveis através da própria plataforma mas também através da aplicação do *Google Earth* ou *Google Maps*. Para serem transferidas para estas aplicações, as fotos passam por um processo de selecção que avalia, segundo as políticas de aceitação, se as fotografias são efectivamente de locais exteriores ou não. Existem excepções relativamente às fotografias de espaços interiores, como por exemplo, se possuem perspectivas de “grande angular” dentro de igrejas, mesquitas, estações de comboios, etc.

A rede incita os utilizadores a captarem informação geográfica, e esta particularidade faz com que o seu conteúdo possua uma componente espacial explícita (Antoniou *et al.*, 2010). Aliás, os mesmos autores, no seu trabalho classificam a informação das redes Flickr e Picassa Web como *spatially implicit*, e das redes Panoramio

---

<sup>5</sup> Consultado em Janeiro, 2016. Em <http://www.panoramio.com/help>

<sup>6</sup> O site oficial da rede é <http://www.panoramio.com/>

e Geograph como *spatially explicit*. Enquanto no primeiro grupo, a informação espacial não é uma das principais motivações, no segundo, é expectável que esse tipo de informação (a localização) seja percebido como um factor motivacional e organizacional.

Por outro lado, sendo esta uma dissertação que aplica métodos de análise espacial, quanto mais exactos (e precisos) forem os dados relacionados com a localização das fotografias, melhores e mais concisos serão os resultados. Neste sentido, uma outra característica que fomentou a escolha desta rede como fonte de dados para o estudo, está intimamente ligada com a exactidão da localização (*positional accuracy*) das suas imagens. Zielstra e Hochmair (2013) analisaram a precisão na localização de 1.433 fotografias procedentes das bases de dados de Flickr e Panoramio. Os autores concluíram que as fotografias disponíveis na segunda rede possuem uma melhor exactidão posicional do que as disponíveis na primeira. No caso das fotografias do Panoramio, para a região da Europa e Asia, determinou-se um erro médio de 0 a 24,5 metros (no caso do Flickr, o erro médio varia de 58,5 a 234,5 metros). Particularmente as fotografias localizadas em áreas urbanas possuíam um erro médio de 0 metros.

No contexto da temática a tratar, os dados do Panoramio evidenciam estar mais alinhados às necessidades de informação, uma vez que procuram representar lugares e espaços abertos da cidade, reduzindo assim, o potencial enviesamento (*bias*) originado pela inclusão na análise, de fotografias de outro tipo de conteúdos (*e.g.* social, comercial, etc.). Além disso, o rigor do posicionamento geográfico assume-se como um aspecto favorável que contribuirá para uma superior qualidade dos resultados.

### **2.3.2.1. Limitações dos dados**

Nas secções anteriores foi discutido o potencial da utilização das *geotagged* fotos provenientes de redes sociais como fontes de dados para a análise do comportamento espaço-temporal dos turistas. No entanto, como referido por Straumann *et al.* (2014), o UGC também apresenta muitos desafios a considerar, tais como a sua falta de estrutura, problemas de qualidade, preconceitos demográficos (sobre/sob representação de certos grupos) e questões de privacidade.

Neste contexto, expõem-se dois aspectos que devem ser considerados quando se analisa este tipo de informação. Em concordância com o afirmado por Straumann *et al.* (2014), os dados apenas são representativos de uma parte da procura real. Apesar de se

constatar uma tendência crescente dos turistas para partilhar as suas experiências de viagens através de redes sociais (Buhalis e Law, 2008; Herrera *et al.* 2015), não se pode garantir que todos eles cumpram o mesmo procedimento. Este facto tem implicações directas na constituição da amostra e, consequentemente, nas inferências ou conclusões geradas a partir dela.

No que diz respeito às características dos turistas que tendem a partilhar as suas experiências. Lo *et al.* (2011), no seu estudo sobre o uso de tecnologias *online* para a partilha de fotos de viagens, determinaram que existe uma forte relação entre a idade, a escolaridade e o rendimento dos indivíduos, e a sua decisão de publicar fotos *online*. São as pessoas mais jovens, com maior nível de escolaridade e com rendimentos mais elevados, as mais propensas a publicar fotos de viagens em redes sociais. Os autores também determinaram que os indivíduos com tendência para partilhar este tipo de fotografias eram viajantes experientes, frequentes (usufruindo regularmente de férias no exterior) e muito motivados por descobrir os destinos num nível mais profundo.

Kádár e Gede (2013), na sua experiência utilizando informação extraída do Flickr concluíram que, nos dados, existe uma maior representação de turistas jovens e de mediana idade do que dos turistas reformados.

Ambos os exemplos mencionados dão uma ideia geral sobre o perfil dos indivíduos (turistas) que tendem a partilhar fotografias das suas viagens em redes sociais. Contudo, não se pretende com isto, afirmar que os utilizadores cobertos pela amostra de dados, neste caso do Panoramio, apresentam as mesmas características. Este aspecto deve ser investigado com mais rigor, através da elaboração de estudos específicos. Sendo este trabalho mais orientado para a análise espacial e temporal dos dados, está fora do seu âmbito tirar conclusões sobre este tema em particular.

Um segundo aspecto está relacionado com a componente temporal dos dados. Quando as fotografias são carregadas na rede, para além da própria imagem e da sua localização, o utilizador pode introduzir informação complementar *e.g.* título, descrição e *tags*, a data em que foi tirada, o modelo da câmara, etc. Todos estes atributos ficam armazenados na base de dados e podem ser visualizados, em conjunto com a imagem, na plataforma web.

No caso do Panoramio, nem toda a informação associada às fotografias se encontra disponível para ser extraída através da API. Relativamente à informação

temporal, só é disponibilizada a data em que a foto foi carregada na rede. Infelizmente, em alguns casos, a data de carregamento pode não coincidir com a data de aquisição e, como tal, não reflectir o momento (real) em que a visita foi realizada.

Tendo-se consciência das implicações do possível desvio nas datas de referência, aquando da análise temporal dos dados, nomeadamente quando se procuram padrões em intervalos de tempo superiores à amplitude dos desvios, baseou-se este trabalho no determinado por Antoniou *et al.* (2010)<sup>7</sup> sobre a diferença aproximada de tempo entre as datas de carregamento e de aquisição das fotos para o caso do Flickr. Assim, a extensão temporal da análise centra-se em períodos mensais e anuais.

Portanto, e tal como referido por Kisilevich *et al.* (2013), o nível de agregação temporal permite mitigar, ou mesmo omitir, o desvio/erro potencialmente presente na referência temporal da informação. A análise espaço-temporal dos dados é apresentada no capítulo 6.

---

<sup>7</sup> Antoniou *et al.* (2010), analisando informação da rede Flickr (esta rede disponibiliza ambas as datas, de carregamento e de aquisição, determinaram que os utilizadores geralmente carregam as suas fotos nas primeiras duas semanas após a data de captura (mais do 60% das fotografias foram carregadas na primeira semana).



### **3. Análise espacial**

#### **3.1. A conjugação da análise espacial e temporal**

Segundo Mitchell (1999), a análise espacial é definida como uma forma de olhar os padrões geográficos dos dados e analisar os relacionamentos entre as entidades. Outros autores definem-na como um conjunto de métodos e técnicas para analisar “eventos” a diferentes escalas, e cujos resultados dependem particularmente da distribuição espacial dos próprios “eventos” (Fisher, 1999); ou como o uso sistemático da localização geográfica dos objectos de interesse como variável importante para a sua descrição, análise e predição (Unwin, 2008).

As técnicas e métodos empregues neste tipo de análise integram, explicitamente, a referência espacial associada a cada um dos dados (Goodchild e Haining, 2004). Numa análise independente do contexto geográfico, toda a informação relevante para a compreensão da variação no conjunto de dados vem simplesmente do atributo da observação, sem considerar nenhuma propriedade referente à sua indexação. No caso dos dados geográficos, a indexação (pelo factor de localização) contém informação essencial para contribuir para a análise (Goodchild e Haining, 2004). Geralmente, mas não necessariamente, a componente espacial é analisada em combinação com os valores dos dados que descrevem os próprios objectos (Unwin, 2008; Fischer, 1999).

Para além da localização espacial, um conjunto de dados pode também incluir, entre os seus atributos, informação alusiva ao momento, *i.e.* tempo em que o levantamento foi efectuado, possibilitando uma indexação espaço-temporal (Andrienko e Andrienko, 2006).

A interacção entre o contexto espacial dos dados e uma referência temporal permite ligar a análise estática de um fenómeno com as suas mudanças ao longo do tempo (Peuquet, 1994). Assim, uma descrição espaço-temporal proporcionará mais informação e, paralelamente, o suporte para uma melhor compreensão dos processos subjacentes à análise de um fenómeno (Illian *et al.*, 2008).

Como já referido, os dados espaciais precisam de técnicas analíticas não convencionais, uma vez que a localização espacial contribui para a sua interpretação (O’Sullivan e Unwin, 2010). Para compreendermos a particularidade desta análise, deve-se rever alguns conceitos de base, apresentados na seguinte secção 3.2.

## 3.2. Conceitos fundamentais para análise espacial

### 3.2.1. Autocorrelação espacial

Um dos principais objectivos da análise de dados é examinar o padrão espacial<sup>8</sup> das variáveis de interesse e perceber a relação entre os valores das entidades próximas (Lloyd, 2010). A autocorrelação espacial refere-se ao facto de que as observações situadas em locais próximos tendem a ser mais semelhantes do que as observações situadas em locais distantes (Getis, 2010).

De forma mais pormenorizada Hubert *et al.* (1981) exemplificam a autocorrelação espacial como “dado um conjunto  $S$  que contém  $n$  unidades geográficas, a autocorrelação espacial refere-se à relação entre uma variável observada em cada um das  $n$  localidades e uma medida de proximidade geográfica definida para todos os  $n(n-1)$  pares seleccionados de  $S$ ”.

Para Miller (2004) a definição de autocorrelação espacial é suportada pelo estabelecido na primeira Lei da Geografia: “*Everything is related to everything else, but near things are more related than distant things*” (Tobler, 1970, p. 274) e, portanto, tem implícito um conceito fundamental: o de proximidade. Tradicionalmente, a proximidade é definida com base no segmento de linha recta que liga duas localizações (distância euclidiana), mas esta é apenas uma das formas de definir “o que está próximo”.

Uma medida de autocorrelação espacial pode sugerir a dependência espacial (autocorrelação espacial positiva - valores semelhantes estão próximos) ou independência espacial (autocorrelação espacial negativa - valores semelhantes não estão próximos). Do mesmo modo, existem outras medidas que descrevem a autocorrelação espacial através das propriedades da distância entre as entidades geográficas (Clark e Evans, 1954; Getis e Ord, 1992).

A maioria dos indicadores conseguem identificar a natureza da autocorrelação e quantificar a sua intensidade, mas, para o efeito, é indispensável estabelecer o conceito de vizinhança (O’Sullivan e Unwin, 2010; Lee e Wong, 2001).

---

<sup>8</sup> Refere-se à “estrutura” de como as observações estão dispostas sobre uma área. O padrão espacial de um fenómeno qualquer varia em função do processo espacial subjacente, podendo ser: concentrado, disperso ou aleatório. (Lee e Wong, 2001).



### 3.2.2. Vizinhança, matriz de proximidade espacial

A matriz de proximidade espacial ou matriz de vizinhança reproduz a nossa compreensão preconcebida sobre as relações espaciais entre as entidades geográficas e representa o conceito de vizinhança a considerar dentro de uma área (Getis, 2010). A análise da estrutura espacial associada a qualquer variável é fortemente dependente da “configuração” da matriz  $W$  de pesos espaciais, onde cada elemento  $W_{ij}$  representa a proximidade entre cada par de unidades espaciais (Rodrigues, 2010). A definição mais simples de vizinhança é expressa através da matriz de contiguidade binária (Waller e Gotway, 2004):

$$w_{ij} = \begin{cases} 1 & \text{se as unidades espaciais } i \text{ e } j \text{ partilham uma fronteira} \\ 0 & \text{caso contrário} \end{cases} \quad (3.1)$$

Onde  $w_{ij}$  é igual a 1 quando  $i$  e  $j$  são contíguos e a 0, caso contrário. Como referido pelos mesmos autores, existem outros esquemas alternativos para quantificar uma relação de proximidade, *e.g.* através do vizinho mais próximo. Neste caso, altera-se a dimensão da vizinhança para abranger uma região próxima, mas não, necessariamente, adjacente. Assim,  $w_{ij}$  toma o valor de 1 se o centroide da região  $j$  é um dos vizinhos mais próximos do centroide da região  $i$  e 0, caso contrário.

Segundo Getis (2010), em vez de se especificar um determinado número de vizinhos mais próximos, pode-se definir a vizinhança através de uma função paramétrica da distância:

$$W_{ij} = d_{ij}^{-\alpha} \quad \alpha \geq 1 \quad (3.2)$$

Onde o peso entre  $i$  e  $j$  é o inverso da distância  $d$  entre as duas localizações,  $i$  e  $j$ , reduzido pelo expoente  $\alpha$ . A matriz  $W$  também pode representar outro tipo de distâncias que não as métricas, *e.g.* distância-tempo, distância-custo. Getis e Aldstadt (2004) e Getis (2010) fazem uma compilação dos esquemas criados para ajustar a matriz  $W$ , *e.g.* comprimento dos limites partilhados dividido pelo perímetro, largura de banda igual a distância entre  $i$  e o  $n$  vizinho mais próximo,  $n$  vizinhos mais próximos, etc.

Neste trabalho, a estrutura da vizinhança é definida com base no declínio da função da distância (inverso da distancia), com um limite estabelecido aos 135m. O procedimento seguido é detalhado na secção 5.3.4.

### 3.2.3. A agregação dos dados

Quando se trata de grandes conjuntos de dados, *i.e. big data*, em muitos os casos, é necessário efectuar um processo de agregação para reduzir a quantidade de dados em análise, *i.e. data reduction*. Este processo agrupa as observações individuais em subconjuntos, cujos atributos associados são gerados a partir dos valores das observações agrupadas (Andrienko e Andrienko, 2006). A agregação de dados, quer a partir de observações individuais para um conjunto de unidades de área, quer com base num mosaico de áreas com geometria específica para outra geometria diferente, pode originar problemas que devem ser tomados em consideração.

A principal questão que envolve a agregação de dados é a perda de informação associada ao processo de passagem de uma representação individual para uma colectiva (Wakefield e Lyons, 2010). São dois os problemas derivados da agregação, a Falácia Ecológica (FA) e a Unidade de Área Modificável (UAM) (Hanning, 2004; Charlton, 2008; Lloyd, 2010; O'Sullivan e Unwin, 2010). O primeiro refere-se à realização de inferências sobre *individuals* suportadas em dados agregados. O *bias* da FA denota às diferenças evidentes entre uma estimativa obtida a partir de dados agrupados e uma estimativa obtida com base em dados individuais (Hanning, 2004). O segundo está relacionado com a variação dos *outputs* devido às mudanças, em dimensão e forma, das áreas de agregação (Lloyd, 2010). O problema da UAM é produzido pelos factores da escala e do zonamento. O efeito da escala postula que os resultados podem estar condicionados pela dimensão da unidade de agregação e, o efeito do zonamento indica que os resultados podem variar em função da forma geométrica de uma área (Charlton, 2008; O'Sullivan e Unwin, 2010; Rodrigues e Tenedório, 2016).

É comum os dados serem agregados, não só para reduzir o volume de informação mas também devido a razões que tornam a questão da agregação indispensável. Neste sentido, a agregação surge também como um procedimento necessário para normalizar a resolução (espacial) dos dados que correspondem a diferentes variáveis (Andrienko e Andrienko, 2006). Contudo, não se deve entender estas características dos dados espaciais como uma limitação para o desenvolvimento de análises.

No âmbito desta dissertação, os dados iniciais (pontos) foram agregados em unidades de área igual a  $2.346\text{m}^2$ . O resultado da contagem das *features* no interior de cada área passou a ser o valor de referência das unidades.

### 3.2.4. Métodos quantitativos de análise espacial

Os métodos quantitativos de análise espacial fornecem um suporte sólido para a avaliação das “ideias” sobre um processo geográfico, podendo ser utilizados para os seguintes fins (Fotheringham *et al.*, 2000):

- Efectuar uma análise exploratória de dados a fim de sugerir hipóteses ou para analisar a presença de *outliers*. Incluem-se também as técnicas de visualização dos dados empregues antes de se efectuar algum outro tipo de análise formal;
- Analisar o papel da aleatoriedade ou “acaso estatístico” na geração do padrão espacial observado e para testar hipóteses sobre o próprio padrão;
- Modelar processos espaciais. Os modelos são utilizados para a geração de valores expectáveis em diferentes cenários, muitas vezes comparados com a realidade.

Segundo Heywood *et al.* (2006), os diferentes métodos quantitativos de análise espacial podem ser enquadrados em três classes: i) Estatísticas exploratórias e descritivas; ii) Estatísticas preditivas; e iii) Estatísticas prescritivas.

Os métodos relacionados com o primeiro tipo centram-se na identificação e descrição dos padrões espaciais ligados ao fenómeno em estudo. As estatísticas exploratórias e descritivas são, geralmente, enquadradas num formato de análise denominado “Análise Exploratória de Dados Espaciais”. Quanto às estatísticas preditivas, são empregues para examinar as relações entre o fenómeno e outras variáveis que contribuem para sua explicação. Por exemplo, através de uma análise de regressão consegue-se perceber a relação entre duas variáveis e calcular (interpolat/extrapolar) valores para áreas omissas nos levantamentos. No que respeita ao terceiro tipo de estatísticas, as prescritivas, estas ajudam com a previsão do que pode acontecer considerando circunstâncias particulares. Aqui, os autores invocam o desenvolvimento de métodos de optimização de soluções (apoiados pela análise de padrões) cujo fim é de dar suporte à tomada de decisões, *i.e.* constituir um sistema de suporte à tomada de decisões (SSTD).

Nesta tese, os métodos foram seleccionados com o objectivo de analisar o padrão espacial do fenómeno em estudo. Portanto, podem ser enquadrados sob a égide dos métodos de tipo um. É também incluído um método para avaliar as relações entre o fenómeno e outras variáveis que contribuiriam para sua explicação.

### **3.3. Análise estatística de dados espaciais**

Na bibliografia de referência podem-se encontrar vários esquemas que conceptualizam o modelo de análise espacial. O'Sullivan e Unwin (2010) expõem que o processo percorre três etapas: Os dados são recolhidos, visualizados e descritos. Em seguida, os dados são explorados para levantar questões e sugerir teorias sobre o fenómeno de interesse. Finalmente, estas teorias são submetidas a testes utilizando técnicas estatísticas espaciais.

Muito semelhante à estrutura anterior e no contexto dos SIG, Charlton (2008) desagrega o mesmo processo em quatro etapas, cada uma das quais podendo envolver várias sub-etapas:

1. Definir o método de análise;
2. Preparação e exploração dos dados (prévia à análise);
3. Análise estatística dos dados;
4. Interpretação dos resultados.

Ambos os esquemas processuais apresentam, claramente, duas fases distintas a seguir quando se efectuar uma análise de dados espaciais. A primeira corresponde à exploração dos dados e a segunda à aplicação de métodos estatísticos em função dos objectivos de trabalho.

#### **3.3.1. Exploração de dados espaciais**

A exploração de dados é uma das etapas fundamentais na análise espacial na medida em que permite ter uma primeira noção das suas características e como tal, do comportamento do fenómeno em estudo. Para Waller e Gotway (2004), a análise exploratória é importante pois fornece informação valiosa que pode orientar a escolha dos métodos estatísticos.

A análise exploratória de dados espaciais (ESDA<sup>9</sup>) precede a aplicação de qualquer método estatístico. Segundo Brunson (2008), a exploração de dados, envolve a execução de um conjunto de “tarefas chave”, entre elas: i) Avaliar a legitimidade dos

---

<sup>9</sup> ESDA é o acrónimo de *Exploratory Spatial Data Analysis* utilizado na bibliografia de referência para denominar ao processo de exploração de dados espaciais (Hanning, 2004; Andrienko e Andrienko, 2006; Longley *et al.*, 2005).

dados; ii) Identificar a presença de registros duvidosos; iii) Identificar a tendência geral nos dados; e iv) Identificar os dados atípicos. As duas primeiras estão destinadas à prevenção de erros nos dados e as subsequentes, à análise inicial (*i.e.* exploratória) dos dados.

Na sua definição de ESDA, Anselin (1994), acrescenta que esta análise é utilizada nomeadamente para visualizar a distribuição espacial dos dados e para avaliar a não estacionariedade de um processo espacial. Na análise exploratória dedica-se muita atenção à representação gráfica dos dados. No caso da informação espacial, a sua visualização não só é possível através de gráficos (como numa análise convencional) mas também através do uso de mapas temáticos que facilitam a interpretação da “distribuição geográfica” dos dados (Brunsdon, 2008).

Em SIG, a visualização dos dados é um processo interactivo. Para representar a informação segundo o que se pretende evidenciar, o ArcGIS disponibiliza alguns métodos de classificação de dados<sup>10</sup>, sendo dois preferencialmente utilizados: Intervalos naturais (*i.e.* diagrama de dispersão) e Desvio padrão.

- Intervalos naturais: esta classificação baseia-se no agrupamento natural inerente aos dados. O conjunto de dados é dividido em classes cujas fronteiras são definidas pela existência de maiores diferenças relativas nos dados. As quebras procuram maximizar as diferenças entre classes e minimizar a variância dentro de cada uma delas.
- Desvio Padrão: o método procura representar quanto os valores de cada uma das entidades se afastam da média. As quebras de classe são criadas em função da proporção dos desvios (0,5; 1,5 e 2,5 desvios padrões relativamente à média).

A visualização através de um mapa expande o quadro de possibilidades para interactivar com a informação. “*The map view of a dataset shows its contents in visual form, and opens many more possibilities for querying*” (Longley *et al.*, 2005, p. 320). Neste contexto, a inquirição (espacial e alfanumérica) torna-se uma operação extremamente útil porquanto facilita a indagação sobre qualquer aspecto dos dados.

---

<sup>10</sup> Em ArcGis Pro *Help*, consultado em Abril, 2016. url: <http://pro.arcgis.com/en/pro-app/help/mapping/symbols-and-styles/data-classification-methods.htm>

A inquirição alfanumérica é realizada a partir de uma expressão SQL<sup>11</sup> que interroga a base de dados associada ao mapa, enquanto a inquirição espacial se baseia em operadores de relações espaciais (*e.g.* topológicas). No *software* a utilizar, estão disponíveis duas ferramentas relacionadas com a inquirição em estruturas vectoriais: Selecção por atributos (alfanumérica) e Selecção por localização (espacial). Ambas permitem seleccionar as *features* que respondem aos critérios especificados.

### 3.3.2. Métodos estatísticos

O processo de análise envolve a aplicação de métodos estatísticos que, ao contrário dos convencionais, integram a referência espacial dos dados como elemento adicional no cálculo (Haining, 2004; Scott e Getis, 2008). Este tipo de estatística baseia-se na não independência das observações, *i.e.* pressupõe que as unidades próximas são de alguma forma associadas (Getis, 1999), o que encontra suporte no postulado na primeira Lei da Geografia.

A estatística espacial compreende um conjunto de técnicas para descrever e modelar dados espaciais (Scott e Janikas, 2010). Nesta secção, são apresentados os métodos utilizados no desenvolvimento do trabalho, tendo sido seleccionados com o objectivo de analisar as características dos dados relativamente à sua distribuição espacial e descobrir os locais atípicos (*outliers*<sup>12</sup>) e/ou de concentração (*clusters* espaciais<sup>13</sup>). Inclui-se também um método para identificar as relações espaciais entre o padrão observado e outros factores explicativos.

Na Tabela 3.1 são descritas, brevemente, as ferramentas seleccionadas para efectuar a análise espacial de dados. É de referir que as ferramentas tem sido citadas segundo a nomenclatura do *software* a utilizar, o ArcGIS® 10.3. (ou o seu equivalente em português). Contudo, deve-se recordar que estas foram desenvolvidas com base em métodos estatísticos, com denominações formais próprias (*e.g.* a ferramenta análise de *clusters* e *outliers* tem o seu suporte no método estatístico denominado *Local Moran's I*). Daqui deriva que, nesta secção, se utilizem os termos “ferramentas de análise” e “métodos

---

<sup>11</sup> SQL ou *Standard query language* é uma linguagem padrão para consultar tabelas ou bancos de dados relacionais (Longley et al., 2005).

<sup>12</sup> Um *outlier* espacial refere-se a uma localidade cujo valor é elevado e está rodeada por valores baixos ou vice-versa (Anselin et al., 2010; Longley et al., 2005; Waller e Gotway, 2004).

<sup>13</sup> Um *cluster* espacial é definido como o excesso de eventos ou de valores num local determinado que difere significativamente da variação geográfica esperada (Jacquez, 2008).

estatísticos” como análogos (apesar das evidentes diferenças respeito ao significado de cada um deles).

**Tabela 3.1: Resumo do conjunto de ferramentas utilizadas na análise dos dados.**

Ferramentas de análise	Descrição
Centro médio	Identifica o centro geográfico do conjunto de <i>features</i>
Elipse do desvio padrão	Mede a concentração das <i>features</i> em torno da média geográfica, e se estas apresentam uma tendência direcional.
<i>Average nearest neighbor</i>	Calcula a distância média entre uma <i>feature</i> e o seu vizinho mais próximo.
<i>Ripley's K-function</i>	Determina a concentração ou dispersão espacial num conjunto de entidades geográficas, em diferentes intervalos de distâncias.
Análise de <i>clusters</i> e <i>outliers</i>	Mede a concentração de valores numa área de estudo.
Análise de pontos quentes emergentes ( <i>Emerging hot spot analysis</i> )	Mede a concentração de valores num eixo tridimensional espaço (X e Y) – tempo (Z).
Regressão Linear Múltipla ( <i>Geographically weighted regression</i> [GWR])	Analisa as relações existentes entre uma variável dependente e múltiplas variáveis independentes (explicativas).

O Centro médio e a Elipse do desvio padrão são ferramentas muito simples, na sua parametrização e no tempo de processamento, cujos resultados sintetizam de forma geral as características da distribuição espacial. As restantes ferramentas, com excepção da GWR, são destinadas à análise de padrões e mapeamento de locais com propriedades relevantes. O processo inicia com a aplicação de ferramentas para determinar a dependência espacial a nível global (tanto o *Average nearest neighbor* como o *K-function* são indicadores de autocorrelação espacial que baseiam o seu cálculo na distância entre as observações), e continua com a aplicação de ferramentas para captar a variação da dependência espacial dentro da área de estudo. De salientar ainda o emprego do *Emerging hot spot analysis* para incluir a componente temporal dos dados na análise de *clusters*.

No contexto dos SIG, as primeiras ferramentas ou funções de análise espacial estão enquadradas na categoria dos “sumários estatísticos” (Matos, 2008) ou, segundo a designação de Longley *et al.* (2005), “sumários descritivos”. Nesta última obra, os autores referem-se ao cálculo de “estatísticas de resumo”, incluindo as medidas que vão desde os equivalentes espaciais das estatísticas univariadas (*e.g.* média, mediana, desvio padrão) às medidas de fragmentação ou de dependência espacial como os indicadores de autocorrelação local.

Quanto à análise temporal mencionada no início deste capítulo, é orientada seguindo duas abordagens. A primeira, através de um processo iterativo que efectua o cálculo para todos os intervalos de tempo estabelecidos na metodologia. Aqui apenas se consideram três das ferramentas indicadas na Tabela 3.1 (o centro médio, a elipse do desvio padrão e o *Ripley's K-function*). A segunda, através da ferramenta *Emerging hot spot analysis*. Neste caso, a diferença para a abordagem anterior é de que a informação temporal de cada uma das observações é inserida directamente no momento da parametrização e como tal, as duas dimensões, espaço e tempo, são analisadas de forma explícita e conjunta, *i.e.* análise espácio-temporal.

Para finalizar, e de forma a explorar as relações entre a variável em estudo e outras consideradas como explicativas, utilizou-se o método de regressão linear múltipla através da ferramenta GWR.

### 3.3.2.1. Características espaciais da distribuição

Estas ferramentas são descritivas por natureza e ajudam a resumir as características mais salientes de uma distribuição espacial (Scott e Janikas, 2010). A partir de um conjunto de dados, é possível representar a tendência da distribuição, relativamente à forma de como os dados estão dispostos no espaço. Neste caso, são utilizadas duas medidas, o centro médio e a elipse do desvio padrão, uma vez que os seus resultados são complementares e sua leitura combinada gera uma noção mais esclarecedora da distribuição.

#### Centro médio e Centro médio ponderado

A forma mais comum de resumir um conjunto de dados é calcular a sua média. O centro médio é uma medida de tendência central, calculada para resumir as posições das entidades a partir dos valores das coordenadas, para cada uma das dimensões X e Y (Fotheringham *et al.*, 2000). Existem duas abordagens para calcular o centro médio, uma que considera apenas a posição das entidades e outra que associa um peso a cada uma das entidades no cálculo, denominada como centro médio ponderado (Mitchell, 2005). O centro médio é dado por:

$$\bar{\chi} = \frac{\sum_{i=1}^n \chi_i}{n} \quad \bar{\gamma} = \frac{\sum_{i=1}^n \gamma_i}{n} \quad (3.3)$$



Onde  $\sum \chi_i$  e  $\sum \gamma_i$  são os somatórios dos valores das coordenadas X e Y das entidades, respectivamente. E  $n$  é o número total de observações. No caso do centro médio ponderado (3.4), a equação é similar a (3.3) mas acrescenta-se o factor ( $w$ ) que corresponde ao peso da entidade  $i$ .

$$\bar{\chi}_w = \frac{\sum_{i=1}^n w_i \chi_i}{\sum_{i=1}^n w_i} \quad \bar{\gamma}_w = \frac{\sum_{i=1}^n w_i \gamma_i}{\sum_{i=1}^n w_i} \quad (3.4)$$

### **Elipse do desvio padrão (Standard Deviatonal Ellipse) e Elipse do desvio padrão ponderada (Weighted Standard Deviatonal Ellipse)**

Outra medida comum para a caracterização da distribuição é produzida com base no cálculo da distância padrão para as coordenadas X e Y face ao centro médio. Cada um dos cálculos para as coordenadas X e Y são efectuados separadamente. Estes valores definem os eixos da elipse que é desenhada à volta do centro médio. A figura da elipse resultante mostra se a distribuição espacial possui alguma orientação em particular. De forma similar à medida anterior (*i.e.* centro médio), a localização das entidades pode ser influenciada por um atributo que actue como peso, neste caso, a medida denomina-se elipse do desvio padrão ponderada (Mitchell, 2005). A equação para o cálculo da elipse do desvio padrão traduz-se por:

$$SDE_{\chi} = \sqrt{\frac{\sum_{i=1}^n (\chi_i - \bar{\chi})^2}{n}} \quad SDE_{\gamma} = \sqrt{\frac{\sum_{i=1}^n (\gamma_i - \bar{\gamma})^2}{n}} \quad (3.5)$$

$$w SDE_{\chi} = \sqrt{\frac{\sum_{i=1}^n w_i (\chi_i - \bar{\chi})^2}{\sum_{i=1}^n w_i}} \quad w SDE_{\gamma} = \sqrt{\frac{\sum_{i=1}^n w_i (\gamma_i - \bar{\gamma})^2}{\sum_{i=1}^n w_i}} \quad (3.6)$$

Onde  $\chi_i$  e  $\gamma_i$  são as coordenadas da entidade  $i$ ,  $\bar{\chi}$  e  $\bar{\gamma}$  correspondem às coordenadas do centro médio do conjunto de entidades e  $n$  é igual ao número total de observações. A equação para o cálculo da elipse do desvio padrão ponderada (3.6) é similar mas incluindo o factor peso. No que respeita ao ângulo de orientação, é calculado através da função trigonométrica (3.7). A orientação representa a rotação do eixo, medido no sentido horário a partir do meio-dia:

$$\begin{aligned}
\tan \theta &= \frac{A + B}{C} \\
A &= \left( \sum_{i=1}^n x_i^2 - \sum_{i=1}^n \tilde{\gamma}_i^2 \right) \\
B &= \sqrt{\left( \sum_{i=1}^n x_i^2 - \sum_{i=1}^n \tilde{\gamma}_i^2 \right)^2 + 4 \left( \sum_{i=1}^n x_i \tilde{\gamma}_i \right)^2} \\
C &= 2 \sum_{i=1}^n x_i \tilde{\gamma}_i
\end{aligned} \tag{3.7}$$

Onde  $x_i$  e  $\tilde{\gamma}_i$  são os desvios das coordenadas XY, com referência ao centro médio.

Por finalizar, as dimensões dos eixos são expressados por:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i \cos \theta - \tilde{\gamma}_i \sin \theta)^2}{n}} \quad \sigma_y = \sqrt{\frac{\sum_{i=1}^n (x_i \sin \theta - \tilde{\gamma}_i \cos \theta)^2}{n}} \tag{3.8}$$

### 3.3.2.2. Índice do vizinho mais próximo

O índice do vizinho mais próximo (*Nearest Neighbor Index* [NNI]) permite constatar se a distribuição espacial das observações é concentrada ou dispersa. Este índice confronta a distância média observada entre cada entidade e o seu vizinho mais próximo e a distância média observada numa distribuição aleatória hipotética, também denominada distância média esperada (Lee e Wong, 2001; Scott e Janikas, 2010). A distância média esperada tem como base uma distribuição aleatória hipotética com o mesmo número de *features* e cobre a mesma área total de estudo.

O NNI relaciona a distância média observada e a distância média esperada, segundo a equação (3.9):

$$\text{NNI} = \frac{\overline{D}_o}{\overline{D}_E} \tag{3.9}$$

Onde  $\overline{D}_o$  é a distância média observada e  $\overline{D}_E$  é a distância média esperada, calculadas com base nas equações (3.10) e (3.11) respectivamente:

$$\overline{D}_o = \frac{\sum_{i=1}^n d_i}{n} \tag{3.10}$$

$$\bar{D}_E = \frac{0,5}{\sqrt{\frac{n}{A}}} \quad (3.11)$$

Nas equações anteriores,  $d_i$  é igual à distância entre uma entidade  $i$  e o seu vizinho mais próximo,  $n$  corresponde ao número total de *features* e  $A$  é a própria área de estudo.

Quando a distância média observada é menor que distância média esperada, o resultado da relação (3.9) é inferior a 1 (indicando uma distribuição concentrada). Quando mais o valor do NNI se aproxima do 0, mais concentrada é a distribuição (Lee e Wong, 2001). No sentido oposto, se a distância média observada é maior, a distribuição é considerada dispersa.

A significância estatística do NNI é aferida através de um teste de hipótese. Para o efeito, a ferramenta calcula um outro índice denominado *z-score* que ajuda a decidir se deveremos rejeitar, ou não, a hipótese nula. Neste caso, a hipótese nula postula que o padrão espacial observado resulta de um processo aleatório e que não é produto de um processo espacial subjacente. O cálculo do *z-score* é expresso da seguinte forma:

$$z = \frac{\bar{D}_o - \bar{D}_E}{SE} \quad (3.12)$$

$$SE = \frac{0,26136}{\sqrt{\frac{n^2}{A}}} \quad (3.13)$$

Onde  $\bar{D}_o$  e  $\bar{D}_E$  correspondem à distância observada e esperada e o *Standard Error (SE)* é o Erro Padrão. Na equação (3.12) compara-se a diferença das distâncias com o erro padrão, o qual se obtém utilizando a equação (3.13).

Os valores de *z-score* estão associados a uma distribuição normal, *i.e.* gaussiana. O resultado do *z-score* pode variar entre valores positivos ou negativos. Para um nível de confiança de 99%, o *z-score* é inferior a -2,58 ou superior a 2,58; para um nível de confiança de 95%, o *z-score* varia entre -2,58 e -1,96 ou entre 1,96 e 2,58 e para um nível de confiança de 90%, o *z-score* varia entre -1,96 e -1,65 ou entre 1,65 e 1,96. O NNI não é estatisticamente significativo se o valor do *z-score* varia entre -1,65 e 1,65 (Mitchell, 2005).

Na avaliação da significância do NNI, também se deve considerar o *p-value*. Este indicador está relacionado com a probabilidade do padrão espacial observado ser criado por algum processo aleatório. Valores muito baixos sugerem uma probabilidade muito reduzida de que o padrão seja resultado de processos aleatórios.

A leitura conjunta do *z-score* e do *p-value* permite determinar a significância estatística do NNI. Neste sentido, ao obter-se valores muito altos ou baixos (valores negativos) no *z-score* e valores muito baixos no *p-value*, pode-se inferir que o padrão espacial observado não reflecte um padrão aleatório (O ArcGis utiliza como hipótese nula a completa aleatoriedade espacial (*Complete Spatial Randomness* [CSR]), quer das próprias *features* quer dos valores que lhes estão associados).

### 3.3.2.3. K-Function

O *K-function* é um método utilizado para identificar propriedades de dependência espacial (concentração ou dispersão) considerando diferentes escalas (Haining, 2004; Dixon, 2002; Scott e Janikas, 2010). Na análise incluem-se todos os vizinhos de cada *feature* ao longo de intervalos de distância, o que permite evidenciar a variação da concentração ou dispersão espacial em função da dimensão da escala de análise (Getis e Franklin, 1987).

Para descrever a estrutura espacial, o método, apoia-se na distância entre pares de eventos e no aumento da escala de análise. O cálculo da função pode-se sintetizar como a razão entre a soma dos pares de *features* dentro de uma distância pré-especificada (de todas as observações) e a soma de todos os pares de *features* tendo como referência a área total de análise (Getis, 1999).

Segundo Getis (1999), a equação original (3.14) do *k-function* introduzida por Ripley em 1977 foi posteriormente modificada por Besag (1977) e por Getis (1984), para considerar a necessidade de estabilizar a variância da função e para incluir os pesos das entidades, respectivamente. A última adaptação, a de 1984, encontra-se disponível no ArcGis (*L-function*) e é utilizada neste estudo (3.15):

$$K(d) = \frac{A}{n^2} \sum_{j \neq i}^n I_{ij} d_{ij} \quad (3.14)$$

$$L(d) = \sqrt{\frac{A \sum_{i=1}^n \sum_{j=1, j \neq i}^n k_{i,j}}{\pi n(n-1)}} \quad (3.15)$$

Onde  $\sum_{i=1}^n \sum_{j=1, j \neq i}^n$  correspondem ao somatório de todas as *features* dentro de uma distância  $d$  da entidade  $i$ , incluindo o *boundary correction*<sup>14</sup> quando requerido.  $n$  é o número total de *features*,  $A$  representa a área total,  $k_{ij}$  corresponde ao peso,  $(n - 1)$  representa todos os pares de eventos possíveis, sendo  $i$  um membro do par; e  $\pi$  é uma área circular centrada na entidade  $i$ . A equação (3.15) multiplica a área pelo número de pares de *features* para uma determinada distância e divide o resultado por  $\pi$  vezes o número possível de pares, sobre a raiz quadrada (Mitchell, 2005).

Como já referido, a  $L(d)$  representa a mesma informação da  $K(d)$ . No entanto, a primeira estabiliza a variância gerada pela “propriedade cumulativa” da segunda ( $K$ ) devido ao incremento do número de eventos com o aumento da distância ( $r$ ), o que facilita também a interpretação do resultado (Illian *et al.*, 2008).

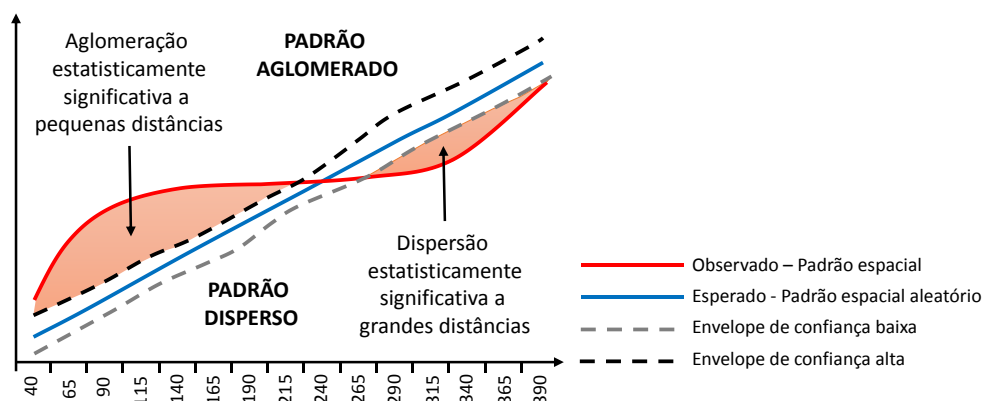
Como parte do procedimento de cálculo, o método começa por traçar círculos concêntricos com um determinado espaçamento e efectua uma contagem das *features* situadas dentro de cada um desses “anéis”. O procedimento é realizado para cada uma das observações e o número acumulado de *features* para cada intervalo é considerado como a função  $K$  (Diggle *et al.*, 1996). Uma vez estimada a função  $K$  para uma distância  $d$ , esta pode ser comparada com o resultado esperado de uma função  $K$  (hipotética) considerando uma distribuição teórica particular, como por exemplo a aleatoriedade no processo espacial (Fotheringham *et al.*, 2000). A comparação permitirá identificar se a distribuição observada é concentrada, ou dispersa, e a correspondente distância (ou a escala) onde estas propriedades são mais acentuadas.

No que respeita ao resultado, a ferramenta *Multi-Distance Spatial Cluster Analysis (Ripley's K-function)* do *software*, gera um gráfico de linhas (Figura 3.1) onde é facilmente interpretável a comparação supramencionada. A linha diagonal azul representa o padrão esperado se as *features* foram distribuídas aleatoriamente dentro da área de estudo. O eixo X mostra os intervalos de distância. A linha curva contínua (vermelha) representa o padrão espacial observado nos dados em análise. Quando a linha vermelha se encontra acima da linha diagonal azul, o padrão é mais concentrado do que o esperado num padrão aleatório, naquela distância específica; quando a linha curva se posiciona

---

<sup>14</sup> O *boundary correction* refere-se à correção feita quando a entidade situa-se nos bordos da área de estudo. A correção baseia-se na assunção de que a área da vizinhança fora da área de estudo apresenta um padrão espacial semelhante à área próxima dentro da área de estudo (Diggle *et al.*, 1996).

abaixo da linha diagonal, o padrão é mais disperso do que o esperado (Scott e Janikas, 2010).



**Figura 3.1: Componentes do gráfico resultante da função K do Ripley.**

**Fonte: Scott e Janikas, 2010.**

A significância estatística para o resultado da função K é avaliada por meio de um teste baseado em simulações (procedimento de Monte Carlo<sup>15</sup>) onde as observações são distribuídas ( $n$  vezes) sob a assunção de CSR (Diggle et al., 1996; Fotheringham et al., 2000; Bivand et al., 2008). Conforme o número especificado de permutações, a ferramenta calcula um envelope de confiança em torno da linha que representa os valores esperados. O número de permutações para o cálculo do envelope está relacionado aos níveis de confiança estatística (O'Sullivan e Unwin, 2010). Por exemplo, para um nível de 90%, 99% ou 99,9% são realizadas 9, 99 ou 999 permutações, respectivamente.

De modo a exemplificar este procedimento, selecionaram-se 99 permutações. A ferramenta vai distribuir aleatoriamente as *features*, 99 vezes para cada iteração. Para cada padrão simulado estima-se os valores de  $k(d)$  e consideram-se os desvios, *i.e.* o valor máximo e o valor mínimo das funções ( $K$ ) para definir as linhas de margem de confiança baixa e alta (Diggle et al., 1996; Dixon, 2000).

Caso a linha dos valores observados esteja entre as linhas de margem de confiança baixa e alta, significa que a concentração ou dispersão não é estatisticamente significativa. Quando a linha dos valores observados se situa fora do envelope de confiança, a concentração ou dispersão é estatisticamente significativa (Scott e Janikas, 2010).

<sup>15</sup> O procedimento de Monte Carlo é qualquer método estatístico que se baseie em numerosas amostragens aleatórias para obter resultados numéricos, *i.e.* repetindo (um elevado número de vezes) simulações sucessivamente, de modo calcular heurísticamente probabilidades, tal como se analisasse os resultados de um jogo de sorte/casino (daí a designação).

### 3.3.2.4. Análise de Clusters e Outliers (ACO)

A identificação dos padrões locais de associação espacial (*Local Indicators of Spatial Association* [LISA]) é fundamental para a análise dos “aspectos espaciais” de um conjunto de dados (Anselin, 1995), uma vez que a intensidade da autocorrelação espacial não é constante ao longo de uma região (Lee e Wong, 2001).

Embora as estatísticas globais possam identificar a presença de uma estrutura espacial (e.g. concentração ou dispersão), estas não conseguem identificar a localização dos *clusters*, nem quantificar a forma como a dependência espacial varia de um lugar para outro (Jacquez, 2008; Scott e Getis, 2008; Lee *et al.*, 2010). Para captar a heterogeneidade da dependência espacial, devem ser aplicados métodos estatísticos direccionados para a escala local.

Em princípio, praticamente qualquer teste geral (estatísticas globais) pode ser transformado numa estatística local (Fotheringham *et al.*, 2000), “*Instead of summarizing over a whole data set, we summarize over only the data in the locality of each data point*” (O’Sullivan e Unwin, 2010, p.222). O LISA refere-se à versão local do *Moran’s I Index*<sup>16</sup> (Anselin, 1995; Anselin *et al.*, 2010) e é expresso por:

$$I_i = \frac{x_i - \bar{X}}{S^2} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X}) \quad (3.16)$$

Onde  $x_i$  é o atributo para a *feature*  $i$ ,  $\bar{X}$  corresponde a média do atributo,  $w_{i,j}$  é o peso espacial (grau de conexão) entre a *feature*  $i$  e  $j$ , e  $S^2$  é calculado por:

$$S^2 = \frac{\sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X})}{n - 1} \quad (3.17)$$

Onde  $n$  é igual ao número total de observações.

O indicador LISA distingue a relação entre o valor de um atributo de uma entidade com os valores correspondentes às entidades na sua vizinhança e consegue identificar a

---

<sup>16</sup> O *Moran I Index* é um indicador para medir o grau de concentração ou dispersão da distribuição espacial de um conjunto de dados, a escala global. Permite identificar se o padrão espacial apresenta uma autocorrelação positiva ou negativa, ou seja, se as entidades próximas (vizinhas) dentro de toda a área de estudo possuem valores semelhantes (autocorrelação espacial positiva) ou valores diferentes (autocorrelação espacial negativa) (Lee e Wong, 2001).

extensão da concentração espacial de valores semelhantes ao redor de cada observação (Anselin, 1995).

Através do método, é possível identificar *clusters*, *i.e.* áreas onde valores elevados estão rodeados por valores elevados ou áreas onde valores baixos estão rodeados por valores baixos. Para além disso, também é possível identificar “irregularidades no padrão” (*outliers*), *i.e.* áreas onde os valores altos são rodeados por valores baixos ou vice-versa (Longley *et al.*, 2005; Waller e Gotway, 2004; Anselin *et al.*, 2010).

A ferramenta *Cluster and Outlier Analysis (Anselin Local Moran's I)* do software ArcGis, também calcula o *z-score* e o *p-value*, como indicadores de suporte à decisão de rejeitar ou não a hipótese nula (Waller e Gotway, 2004). Similarmente ao já mencionado relativamente às outras estatísticas, a obtenção apenas dos valores do *Local Moran's I* para cada observação não assegura a sua significância estatística, uma vez que os resultados do *I* (altos ou baixos) podem ocorrer por acaso, *i.e.* de forma aleatória. Estes valores têm que ser comparados com os valores esperados e interpretados através dos seus *scores* padronizados, *i.e.* *z-score* (Lee e Wong, 2001).

Também é calculado o *p-value* que se refere à probabilidade de que os valores observados em cada LISA superem os valores esperados, sob a premissa da hipótese nula (Waller e Gotway, 2004). A hipótese nula afirma que os valores associados às *features* são distribuídos aleatoriamente e não por um processo subjacente. Neste caso, a hipótese nula diferencia-se ligeiramente das referidas noutros métodos, uma vez que a conceito de aleatoriedade espacial incide sobre a distribuição dos valores da variável e não sobre a localização *per se* das entidades.

Para cada uma das observações, a ferramenta gera o resultado do índice local de Moran, o *z-score*, o *p-value* e um código denominado “COType” que identifica, no mapa, se a *feature* é considerada como *cluster* ou *outlier*.

No que respeita ao índice local de Moran, se o seu valor é positivo indica a concentração de valores numa mesma área. Existem dois tipos de *clusters*: *High-High* (HH) e *Low-Low* (LL), o primeiro representa a concentração de valores elevados e, o segundo, a concentração de valores baixos. Se o índice for negativo, indica a presença de *outliers*. Existem também dois tipos de *outliers*: *High-Low* (HL), preconizando um valor elevado que na sua vizinhança possui principalmente valores baixos e, *Low-High* (LH),



onde um valor baixo se encontra rodeado principalmente por valores elevados (O'Sullivan e Unwin, 2010; Anselin *et al.*, 2010).

De modo a facilitar a interpretação dos resultados, a ferramenta cria uma nova camada de informação (*layer*) onde cada uma das observações são classificadas de acordo com o valor do índice e a sua significância estatística. Por defeito é considerado um nível de confiança de 95%, *i.e.* a hipótese nula é rejeitada unicamente se a probabilidade de que o padrão seja criado por acaso for inferior a 5%. Assim, as *features* são classificadas como HH, HL, LH, LL ou não significativas.

### **3.3.2.5. *Análise de pontos quentes emergentes (Emerging hot spot analysis)***

Esta ferramenta, inserida no módulo *Space time pattern mining tools* do ArcGis, serve para identificar tendências em séries de dados espaço-temporais. Pretende encontrar, por exemplo, pontos estatisticamente significativos de valores elevados (*hot spots*) ou de valores baixos (*cold spots*) novos, em intensificação, em diminuição ou esporádicos.

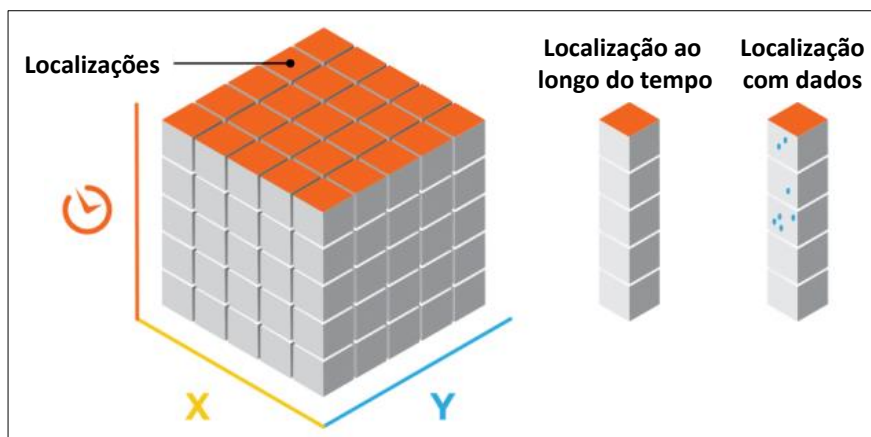
Como *input*, a ferramenta requer que os dados sejam agregados e estruturados num cubo de dados. No próximo ponto explica-se o processo de agregação, sendo a base do método discutida no seu seguimento.

### **Agregação espaço-temporal dos dados num cubo de dados**

O cubo de dados é criado através da ferramenta *Create Space Time Cube*, inserida no mesmo módulo. Assim, agregam-se as *timestamped features* (pontos) em caixas espaço-temporais que, em conjunto, conformam um cubo de dados (seguindo um formato *Network Common Data Form* [NetCDF]<sup>17</sup>). Dentro de cada caixa, os pontos são contados (caso necessário, podem-se apurar estatísticas de resumo dos atributos associados às entidades), e a tendência dos valores ao longo do tempo, em cada local, é medida utilizando a estatística de Mann-Kendall. A estrutura do cubo tem linhas, colunas e passos de tempo (Figura 3.2). As linhas e colunas (eixos X e Y) determinam a extensão espacial do cubo, enquanto os passos de tempo (eixo Z) determinam a extensão temporal. Multiplicando o número de linhas, pelo número de colunas e pelo número de passos do tempo, ter-se-á o número total de compartimentos do cubo.

---

<sup>17</sup> NetCDF é um formato de arquivo para armazenar dados multidimensionais, tais como temperatura, humidade, velocidade e direcção do vento.



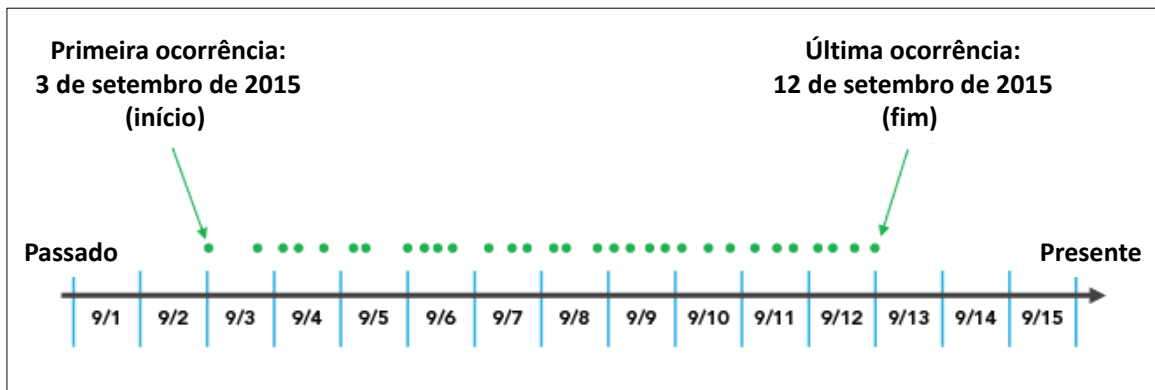
**Figura 3.2: Estrutura do cubo espaço-temporal.**  
**Fonte: Adaptado de ArcGis Help.**

### **Parâmetros de base para a geração do cubo**

Para a criação do cubo, deve-se definir: i) a dimensão, ou seja, a extensão espacial e temporal das caixas de agregação; e ii) o alinhamento espaço-temporal. A dimensão das caixas precisa de ser adaptada às questões particulares que se pretende responder. Analisando fotografias de visitantes, por exemplo, podem-se juntar os pontos em caixas de agregação de 400m<sup>2</sup> por essa ser a dimensão média de um quarteirão na cidade. Quando os dados cobrem a totalidade do ano, pode-se optar por analisar as tendências em termos de agregação de eventos mensais ou semanais.

O alinhamento espaço-temporal determina onde começa e termina a agregação dos dados. Para este parâmetro, a ferramenta disponibiliza três opções: o alinhamento das observações conforme à data de término (data do registro mais recente), ou de início (data do registro mais antigo), ou conforme um tempo de referência específico.

Na Figura 3.3, pode-se observar um conjunto de dados (englobando o período de 3 de Setembro de 2015 a 12 de Setembro de 2015) que será utilizado para exemplificar e explorar as implicações das diferentes opções disponíveis para o alinhamento temporal dos dados.



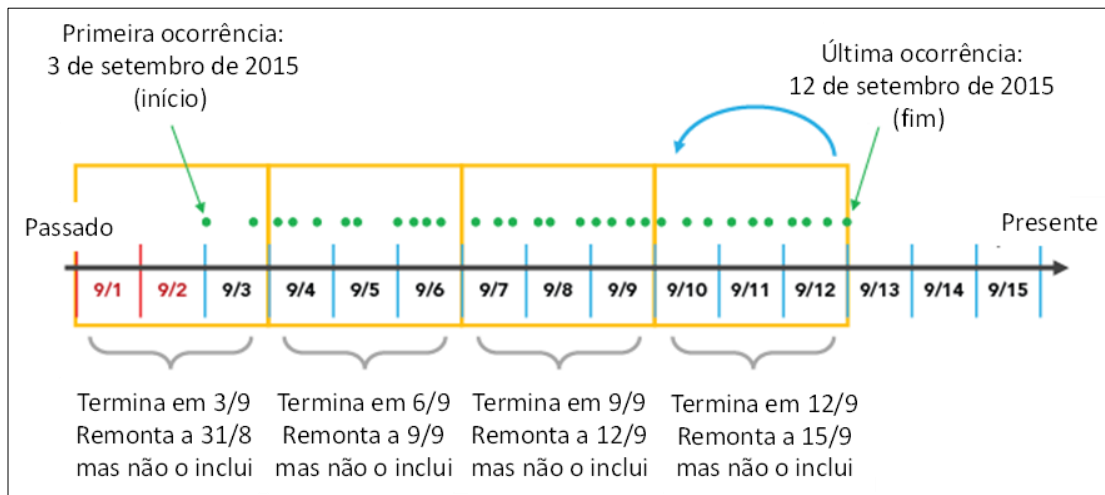
**Figura 3.3: Conjunto de dados abrangendo o período 3 de Setembro de 2015, a 12 de Setembro de 2015. Fonte: Adaptado de ArcGis Help.**

a) No caso de estabelecer o alinhamento conforme a data de término:

Se a data de término do alinhamento espaço-temporal é escolhida com um intervalo de tempo, por exemplo, de 3 dias, a agregação irá ter início na última observação, retrocedendo em incrementos de 3 dias até que todas as ocorrências estejam inseridas num intervalo.

É importante observar que, dependendo do intervalo de tempo escolhido, é possível criar um incremento temporal no início do cubo de espaço-tempo onde não se observam ocorrências ao longo de toda a extensão temporal. Na Figura 3.4, se pode verificar que os dias 1 e 2 de Setembro estão incluídos no primeiro passo de tempo, apesar de existirem dados só até o 3 de Setembro. Estes dias vazios fazem parte do passo de tempo, mas não têm nenhum dado associado. Isto pode influenciar os resultados porque parecerá que o passo temporalmente enviesado tem significativamente menos pontos do que os outros passos de tempo, o que na verdade é um resultado artificial do regime de agregação escolhido. Neste caso, 2 dias (em 3) do primeiro passo não possuem nenhum dado, pelo que o enviesamento temporal seria de 66%.

A data de término é a opção padrão para alinhamento do passo de tempo porque muitas análises concentram-se sobre o que aconteceu mais recentemente. Outra solução, que resolve a questão da tendência temporal como um todo, seria fornecer dados que são divididos igualmente pelo intervalo de tempo, para que não haja períodos tendenciosos.



**Figura 3.4: Alinhamento espaço temporal por data de término.**

**Fonte: Adaptado de ArcGis Help.**

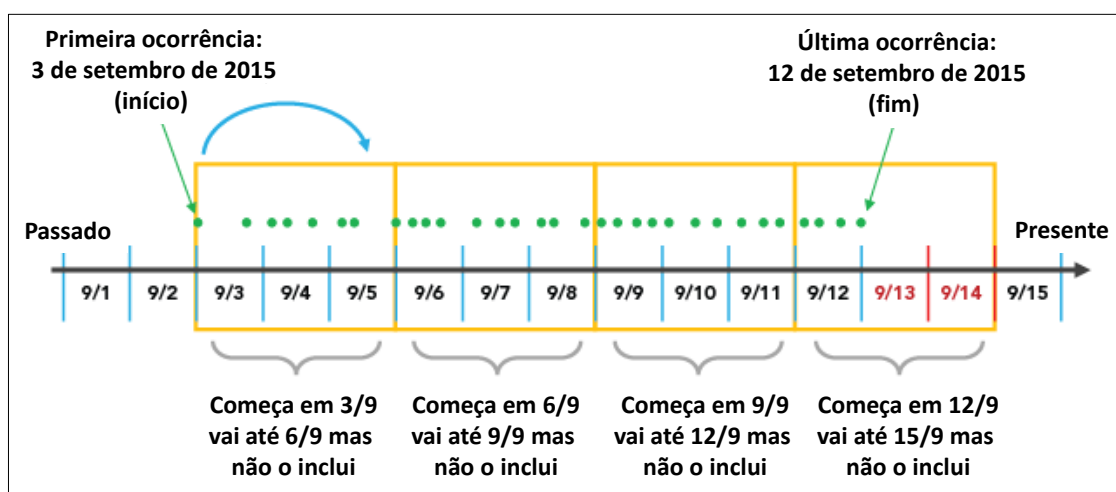
Isto pode ser feito através da criação de um conjunto de dados seleccionados que exclui a parte que fica fora do que se pretendia que fosse o primeiro período de tempo. Assim, seleccionar-se-iam todos os dados, excepto aqueles que ocorrem antes de 4 de Setembro. Neste caso, o intervalo de tempo da primeira e última etapa é usado para determinar onde fazer o “corte”.

Também é importante notar que, se no processo de retroceder no tempo, a caixa final cair exactamente sobre a primeira ocorrência (dada como início do conjunto de dados), essa ocorrência não será incluída na caixa de agregação. Isto, porque cada caixa inclui a última data inserida no seu domínio, mas ao retroceder não inclui a primeira data ocorrida nessa caixa. Nesse caso, teria que ser criada uma caixa adicional para garantir que os dados da primeira ocorrência fossem incluídos.

b) No caso de estabelecer o alinhamento conforme a data de início:

Se for escolhida uma data de início para o alinhamento do passo tempo, seguindo o mesmo exemplo, com um intervalo de tempo de 3 dias, então a caixa vai ter início no primeiro ponto de dados e apresentar incrementos de 3 dias até que o último dado seja inscrito no passo do tempo final (Figura 3.5).

A semelhança do caso anterior, quando escolhe-se esta opção, é possível que o período final não tenha dados em toda a sua extensão temporal. Na Figura 3.5, se pode observar que os dias 13 e 14 de Setembro estão incluídos no último passo de tempo, apesar de não existir nenhuma ocorrência posterior a 12 de Setembro. Estes dias fazem parte do passo tempo mas não têm dados associados. Isto pode influenciar os resultados porque poderá parecer que o último passo de tempo (temporalmente tendencioso) tem significativamente menos pontos do que os outros períodos, o que na verdade é um resultado artificial do regime de agregação. Poderá existir um enviesamento temporal no último passo.



**Figura 3.5: Alinhamento espaço-temporal por data de início.**

**Fonte: Adaptado de ArcGis Help.**

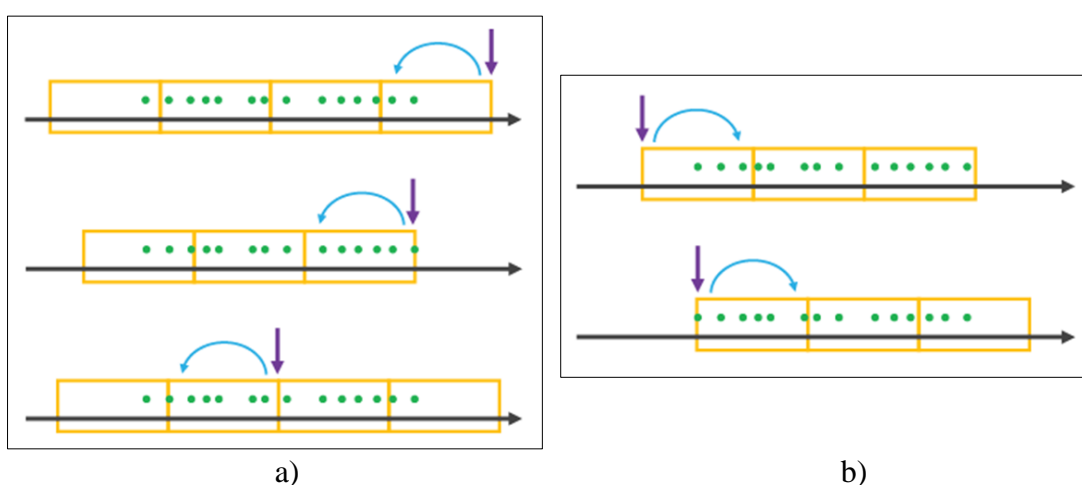
A solução será utilizar dados divididos equitativamente pelos intervalos de tempo, para que não haja períodos tendenciosos. Também se deve notar que, à imagem do que ocorre quando se define uma data de término, se no processo de se avançar no tempo, o fim do último período coincidir exactamente com a última ocorrência, essa ocorrência não será incluída na respectiva caixa de agregação. Isto deve-se ao facto de que com uma data de início para o alinhamento de intervalos de tempo, cada caixa de agregação inclui o primeiro registo mas não inclui a última data a cair dentro da sua área. Então, teria que ser adicionada uma caixa de agregação adicional para assegurar a inclusão da última observação.

c) No caso de estabelecer o alinhamento conforme um tempo de referência

O tempo de referência permite que uma determinada data seja marcada como o início ou o fim de um dos períodos no cubo. Ao escolher um tempo de referência que cai

após a extensão do conjunto de dados, no último ponto ou no meio do conjunto de dados, este será tratado como o último ponto de dados de um intervalo de tempo, e todas as outras caixas de ambos os lados serão criadas usando o alinhamento passo tempo especificado, até que todos os dados sejam cobertos (Figura 3.6 a).

Optando por um tempo de referência que cai antes da extensão do conjunto de dados, ou no primeiro ponto dos dados, este será tratado como o primeiro ponto de um intervalo de tempo, e todos os outros passos de tempo, serão criados usando esta data de início até que todos os dados sejam considerados (Figura 3.6 b).

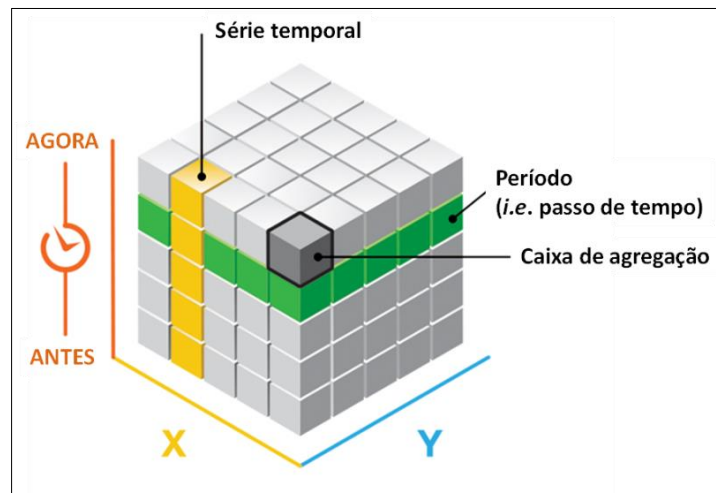


**Figura 3.6: Alinhamento espaço-temporal segundo um “Tempo de referência”.**  
**Fonte: Adaptado de ArcGis Help.**

Ao escolher um tempo de referência fixado antes ou após a extensão dos dados cria-se o potencial para gerar caixas de agregação totalmente ou parcialmente vazias, o que pode influenciar a análise.

A ferramenta também avalia a tendência dos dados agregados, efectuando uma análise de séries temporais espaciais. No fundo, pretende responder à questão de se os eventos representados pelos dados de entrada vão aumentando ou diminuindo ao longo do tempo. Para obter a resposta, o número de pontos para todos os locais em cada intervalo de tempo é analisado através do teste de Mann-Kendall (MK).

O teste de tendência de MK é executado em todos os locais com dados. A estatística de MK é uma análise de correlação do valor (ou contagem) das ocorrências nas caixas de agregação e da respectiva série temporal (Figura 3.7).



**Figura 3.7: Caixas de agregação e relação espaço-tempo.**  
**Fonte: Adaptado de ArcGis Help.**

O valor da caixa de agregação para o primeiro período é comparado com o valor para o segundo período. Se o primeiro for menor que o segundo, o resultado é + 1. Se o primeiro for maior do que o segundo, o resultado é -1. Se os dois valores forem iguais, o resultado é zero. Logo, procede-se ao somatório do resultado de cada par - de passos de tempo - comparado. A soma prevista é zero, não indicando nenhuma tendência dos valores ao longo do tempo. Isto é, com base na variância para os valores da caixa de agregação da série temporal, o número de igualdades e o número de períodos, a soma observada é comparada à soma esperada (zero) para determinar se a diferença é estatisticamente significativa ou não. A tendência para cada caixa de agregação com dados da série de tempo, é assinalada com um *z-score* e um *p-value*. O sinal associado ao *z-score* determina se a tendência corresponde a um aumento nos valores da caixa de agregação (*z-score* positivo) ou uma diminuição desses mesmos valores (*z-score* negativo).

### **Teste de Mann-Kendall**

O objectivo do teste de MK (Mann, 1945; Kendall, 1975; Gilbert, 1987) é avaliar estatisticamente se existe uma tendência monotónica (*i.e.* que se repete ou não varia), ascendente ou descendente, da variável estudada ao longo do tempo. A tendência monotónica para cima (ou para baixo) significa que o valor da variável aumenta (ou diminui) consistentemente ao longo do tempo, mas esta tendência pode ou não ser linear. O teste de MK pode ser usado em substituição de uma análise de regressão linear paramétrica, que pode ser usada para testar se o declive estimado da reta de regressão

linear é diferente de zero. A análise de regressão requer que os resíduos da linha de regressão sejam uma distribuição normal, *i.e.* gaussiana. Este requerimento não é necessária para o teste MK, *i.e.* o MK é um teste não-paramétrico (de distribuição livre).

Hirsch *et al.* (1982) indicam que o teste MK é muitas vezes visto como uma análise exploratória mas é usado com mais propriedade para identificar os pontos onde as mudanças são significativas ou de grande magnitude, e para quantificar esses achados. Assim, o teste de MK original procura uma tendência numa série temporal, sem especificar se a tendência é linear ou não linear (Khaliq *et al.*, 2009). Baseia-se na hipótese nula de que dados são independentes e ordenados aleatoriamente, *i.e.* que não há nenhuma tendência ou estrutura de correlação em série, presente nas observações. O teste MK é descrito por Yue *et al.* (2002) e baseia-se nas estatísticas de teste  $S$ , sendo definido como:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i) \quad (3.18)$$

onde,  $x_j$  são os valores sequencias dos dados,  $n$  é a dimensão do conjunto de dados e,

$$\text{sgn}(t) = \begin{cases} 1, & \text{for } t > 0 \\ 0, & \text{for } t = 0 \\ -1, & \text{for } t < 0 \end{cases} \quad (3.19)$$

O valor de  $S$  indica a direcção da tendência. Um valor negativo (positivo) indica uma tendência decrescente (crescente). Mann-Kendall escreveram que quando  $n \geq 8$ , a estatística de teste  $S$  aproxima-se de uma distribuição normal, com a média e a variância calculadas segundo (1.3) e (1.4), respectivamente:

$$E(S) = 0 \quad (3.20)$$

$$\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^m t_i(t_i-1)(2t_i+5)}{18} \quad (3.21)$$

Onde,  $m$  é o número de grupos vinculados e  $t_i$  é a dimensão do  $i$ -ésimo grupo criado. A estatística de teste padronizado  $Z$  é calculada da seguinte forma:



$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, & \text{for } S > 0 \\ 0 & \text{for } S = 0 \\ \frac{S+1}{\sqrt{Var(S)}}, & \text{for } S < 0 \end{cases} \quad (3.22)$$

A estatística normalizada de Mann-Kendall Z segue a distribuição normal padrão com zero de média e variância igual à unidade. Se  $|Z| \geq Z_{1-(\alpha/2)}$ , a hipótese nula sobre a inexistência de tendência é rejeitada no nível  $\alpha$  de significância.

Para o teste original de MK, a série temporal deve ser independente em série. No entanto, em muitas situações os dados reais e observados são autocorrelacionados. A autocorrelação nos dados observados pode causar má interpretação da tendência dos resultados de teste. Cox e Stuart (1955) afirmaram que uma correlação em série positiva entre as observações aumentaria a probabilidade de uma resposta significativa, mesmo na ausência de uma tendência. Um problema intimamente relacionado e que tem sido estudado é o caso onde existe sazonalidade nos dados (Hirsch *et al.*, 1982).

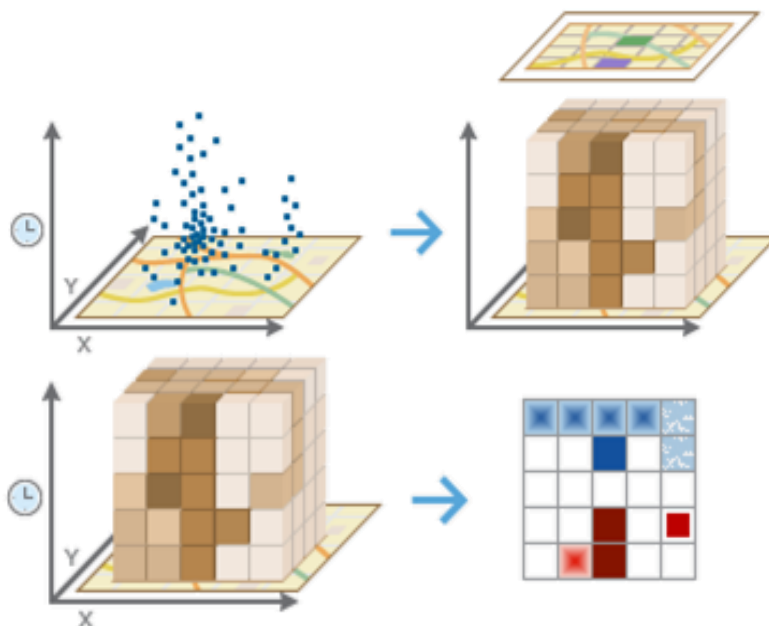
Ao dividir as observações em classes separadas de acordo com a época e em seguida, executar o teste de tendência de Mann-Kendall para a soma das estatísticas de cada período, o efeito da sazonalidade pode ser eliminado. Esta modificação é conhecida como teste de Mann-Kendall sazonal (Hirsch *et al.*, 1982; Hirsch e Slack, 1984). Embora o teste sazonal elimine o efeito da dependência sazonal, isso não é tido em conta para a correlação da série dentro do período (Hirsch e Folga, 1984). O mesmo problema existe quando são consideradas séries anuais como período de análise, pois estas podem ser significativamente autocorrelacionadas.

Como *outputs* desta ferramenta, obtém-se um ficheiro com a representação NetCDF dos dados, acompanhada de um relatório que sumariza as características do cubo (número de caixas de agregação; número de colunas e filas; informação sobre a tendência temporal dos dados, com respectivo teste de significância).

### **Análise de pontos quentes emergentes (APQE)**

O método tem como base a estatística de Getis-Ord Gi (*hot spot analysis*). Na Figura 3.8, tenta-se representar o fluxo de processamento de dados da análise de pontos quentes emergentes. Para cada caixa de agregação do cubo, a ferramenta calcula a

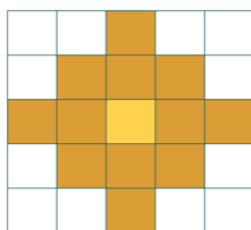
estatística de *Getis-Ord Gi*, considerando o valor da própria caixa (*target*) e das caixas na sua vizinhança (estabelecida através dos parâmetros de distância espacial e temporal).



**Figura 3.8: Fluxo de processamento de dados da análise de pontos quentes emergentes.**  
**Fonte: Adaptado de ArcGis Help.**

Para determinar se o valor de contagem de uma caixa de agregação situada num ponto no espaço e no tempo constitui um ponto quente ou frio estatisticamente significativo, cada caixa de agregação é avaliada dentro do contexto das suas vizinhas espaciais e temporais. Assim, uma caixa de agregação com valor elevado será um ponto quente estatisticamente significativo, a menos que as suas vizinhas no espaço-tempo também apresentem valores elevados.

Os parâmetros para a distância de vizinhança espacial e vizinhança temporal definem a extensão da vizinhança de cada caixa de agregação (*i.e.* o contexto de análise de cada caixa). Supondo que a caixa de agregação possuía as dimensões de 400m<sup>2</sup> por 1 dia, se a distância de vizinhança for fixada em 801 metros e a vizinhança temporal elevada para 2 dias, os vizinhos espaciais vão estender-se duas caixas, tanto horizontal como verticalmente, e uma caixa diagonalmente, correspondendo a uma janela circular de vizinhança de Von Newman (*i.e.* Kings case) com uma dimensão de 5x5 caixas (Figura 3.9).



**Figura 3.9: Vizinhança de Von Newman (5x5).** Fonte: Adaptado de ArcGis Help.

Complementarmente, também existem vizinhos temporais. Todas as caixas de agregação com a mesma localização geográfica da caixa central (em análise) e dos seus vizinhos espaciais (Figura 3.9) para a correspondência do período - num total de três dias para este exemplo - serão incluídas como vizinhas. Observe-se que vizinhos temporais estão para trás no tempo e que uma vizinhança de passo de tempo de 2 engloba três intervalos de tempo.

Uma vez concluída a análise de pontos quentes no espaço-tempo, cada caixa de agregação do cubo (NetCDF) de entrada vai passar a ter uma classificação de “pontos quentes”, um *z-score* e um *p-value*. Em seguida, estas tendências de pontos quentes e frios são avaliadas usando o teste de tendência Mann-Kendall. Com a tendência resultante do *z-score* e do *p-value* para cada local com dados do teste de MK (do cubo de dados), e com o *z-score* e *p-value* para cada caixa de agregação, a ferramenta *Emerging Hot Spot Analysis* classifica cada local da área de estudo como indicado na Tabela 3.2.

**Tabela 3.2. Classes de pontos quentes e frios emergentes.** Adaptado de ArcGIS Help.

Designação	Definição
Nenhum padrão detectado ( <i>No Pattern Detected</i> )	Não se enquadra em nenhum dos padrões de pontos quentes ou frios, definidos a seguir.
Novo ponto quente ( <i>New Hot Spot</i> )	Um local que é um ponto quente estatisticamente significativo no último passo de tempo mas nunca foi um ponto quente estatisticamente significativo antes.
Ponto quente consecutivo ( <i>Consecutive Hot Spot</i> )	Um local com um único e ininterrupto conjunto de caixas de agregação correspondentes a pontos quentes estatisticamente significativos nos intervalos de tempo finais. O local nunca foi um ponto quente estatisticamente significativo antes do último período.
Ponto quente em acréscimo ( <i>Intensifying Hot Spot</i> )	Um local que tem sido um ponto quente estatisticamente significativo em 90% dos intervalos de tempo, incluindo o período final. Além disso, a intensidade de agregação (de valores elevados) em cada passo de tempo, em geral, vai aumentado, e esse aumento é estatisticamente significativo.
Ponto quente persistente ( <i>Persistent Hot Spot</i> )	Um local que tem sido um ponto quente estatisticamente significativo em 90% dos intervalos de tempo mas com nenhuma tendência discernível, indicando um aumento ou diminuição da intensidade do <i>cluster</i> ao longo do tempo.

<b>Designação</b>	<b>Definição</b>
Ponto quente em decréscimo ( <i>Diminishing Hot Spot</i> )	Um local que tem sido um ponto quente estatisticamente significativo em 90% dos intervalos de tempo, incluindo o passo do tempo final. Além disso, a intensidade dos <i>clusters</i> (de valores elevados) em cada passo de tempo está, em geral, a diminuir, e essa diminuição é estatisticamente significativa.
Ponto quente esporádico ( <i>Sporadic Hot Spot</i> )	Um local que consecutivamente é, e deixa de ser, ponto quente. Menos de 90% dos intervalos de tempo são pontos quentes estatisticamente significativos e nenhum dos intervalos de tempo foram pontos frios estatisticamente significativos.
Ponto quente oscilante ( <i>Oscillating Hot Spot</i> )	Um ponto quente estatisticamente significativo para o intervalo de tempo final que também foi um ponto frio estatisticamente significativo durante um período prévio. Menos de 90% dos intervalos de tempo foram pontos quentes estatisticamente significativos.
Ponto quente histórico ( <i>Historical Hot Spot</i> )	O período mais recente não é um ponto quente mas pelo menos 90% dos intervalos de tempo foram pontos quentes estatisticamente significativos.
Novo ponto frio ( <i>New Cold Spot</i> )	Um local que é um ponto frio estatisticamente significativo no passo de tempo final mas nunca antes foi um ponto frio estatisticamente significativo.
Ponto frio consecutivo ( <i>Consecutive Cold Spot</i> )	Um local com um único e ininterrupto conjunto de caixas de agregação correspondentes a pontos frios estatisticamente significativos nos intervalos de tempo finais. O local nunca foi um ponto frio estatisticamente significativo antes do último período.
Ponto frio em acréscimo ( <i>Intensifying Cold Spot</i> )	Um local que tem sido um ponto frio estatisticamente significativo em 90% dos intervalos de tempo, incluindo o período final. Além disso, a intensidade de agregação (de valores baixos) em cada passo de tempo, em geral, vai aumentando, e esse aumento é estatisticamente significativo.
Ponto frio persistente ( <i>Persistent Cold Spot</i> )	Um local que tem sido um ponto frio estatisticamente significativo em 90% dos intervalos de tempo mas com nenhuma tendência discernível, indicando um aumento ou diminuição da intensidade do <i>clustering</i> ao longo do tempo.
Ponto frio em decréscimo ( <i>Diminishing Cold Spot</i> )	Um local que tem sido um ponto frio estatisticamente significativo em 90% dos intervalos de tempo, incluindo o passo do tempo final. Além disso, a intensidade dos <i>clusters</i> (de valores baixos) em cada passo de tempo está, em geral, a diminuir, e esse decréscimo é estatisticamente significativo.
Ponto frio esporádico ( <i>Sporadic Cold Spot</i> )	Um local que consecutivamente é, e deixa de ser, ponto frio. Menos de 90% dos intervalos de tempo são pontos frios estatisticamente significativos e nenhum dos intervalos de tempo foram pontos quentes estatisticamente significativos.
Ponto frio oscilante ( <i>Oscillating Cold Spot</i> )	Um ponto frio estatisticamente significativo para o intervalo de tempo final que também foi um ponto quente estatisticamente significativo durante um período prévio. Menos de 90% dos intervalos de tempo foram pontos frios estatisticamente significativos.
Ponto frio histórico ( <i>Historical Cold Spot</i> )	O período mais recente não é um ponto frio, mas pelo menos 90% dos intervalos de tempo foram pontos frios estatisticamente significativos.

### 3.3.2.6. *Regressão Linear Múltipla*

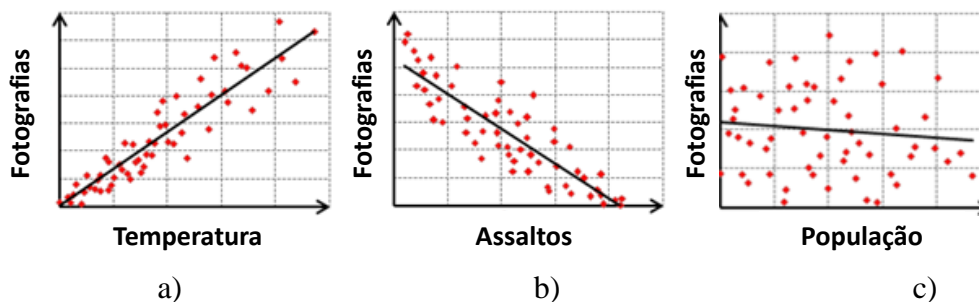
A análise de regressão permite modelar, examinar e explorar as relações espaciais, podendo ajudar a explicar os factores por detrás dos padrões espaciais observados. Pode-se querer perceber as razões porque as pessoas persistentemente tiram fotografias em certas áreas da cidade ou que factores contribuem para uma taxa de fotografias superior à esperada. No entanto, ao modelar as relações espaciais, a análise de regressão também pode ser utilizada para predição.

A análise dos mínimos quadrados (*Ordinary Least Squares* [OLS]) é a mais conhecida de todas as técnicas de regressão. É também o ponto de partida adequado para todas as análises de regressão espacial. Esta técnica fornece um modelo global da variável ou processo que se está a tentar compreender ou prever; cria uma equação de regressão simples para representar esse processo. No seguimento da OLS, a regressão geográfica ponderada (*Geographically Weighted Regression* [GWR]) é uma das várias técnicas de regressão especial que cada vez mais é utilizada em Geografia.

A GWR providencia um modelo local da variável ou processo a modelar, através do ajuste de uma equação de regressão a cada elemento (factor) considerado. Quando usado correctamente, este método fornece estatísticas importantes - e de confiança - para examinar e estimar relações lineares. Estas podem ser positivas ou negativas. Por exemplo, comprovando que o número de fotografias aumenta quando as temperaturas diurnas sobem, obtém-se uma relação dita positiva, *i.e.*, existe uma correlação positiva. Outra forma de expressar essa relação positiva é dizer que os eventos captação de fotografias decrescem à medida que as temperaturas diurnas diminuem.

Por outro lado, se o número de fotografias diminuir face ao aumento de assaltos por parte de carteiristas, a relação é negativa. A relação negativa pode ser expressa afirmando que o número fotografias aumenta à medida que o número de assaltos diminui. A Figura 3.10 retrata as relações positivas e negativas, bem como o caso onde não há nenhuma relação entre duas variáveis.

As análises de correlação e os gráficos associados retratados na Figura 3.10, testam a força do relacionamento entre duas variáveis. As análises de regressão, por outro lado, fazem uma reivindicação mais forte pois tentam demonstrar o grau segundo o qual, uma ou mais variáveis, potencialmente promovem uma mudança, positiva ou negativa, numa outra variável.



**Figura 3.10: Diagramas de dispersão: relação positiva (a), relação negativa (b), e duas variáveis não relacionadas (c)**

A equação de regressão aplicada às variáveis explicativas tidas como as que melhor ajudam a prever o comportamento da variável dependente que se pretende modelar, pode ser expressa como:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (3.23)$$

Onde  $Y$  é a variável dependente, os  $X$  são as variáveis explicativas e os  $\beta$  são coeficientes de regressão. Cada variável independente está associada a um coeficiente de regressão, o qual descreve a força e o sinal da relação dessa variável com a variável dependente. Infelizmente para quem trabalha em Ciência de Informação geográfica, e que pensa em  $x$  e  $y$  como coordenadas, a notação em equações de regressão para a variável dependente é sempre  $Y$  e para as variáveis independentes (ou explicativas) é sempre  $X$ , o que pode conduzir a alguns constrangimentos.

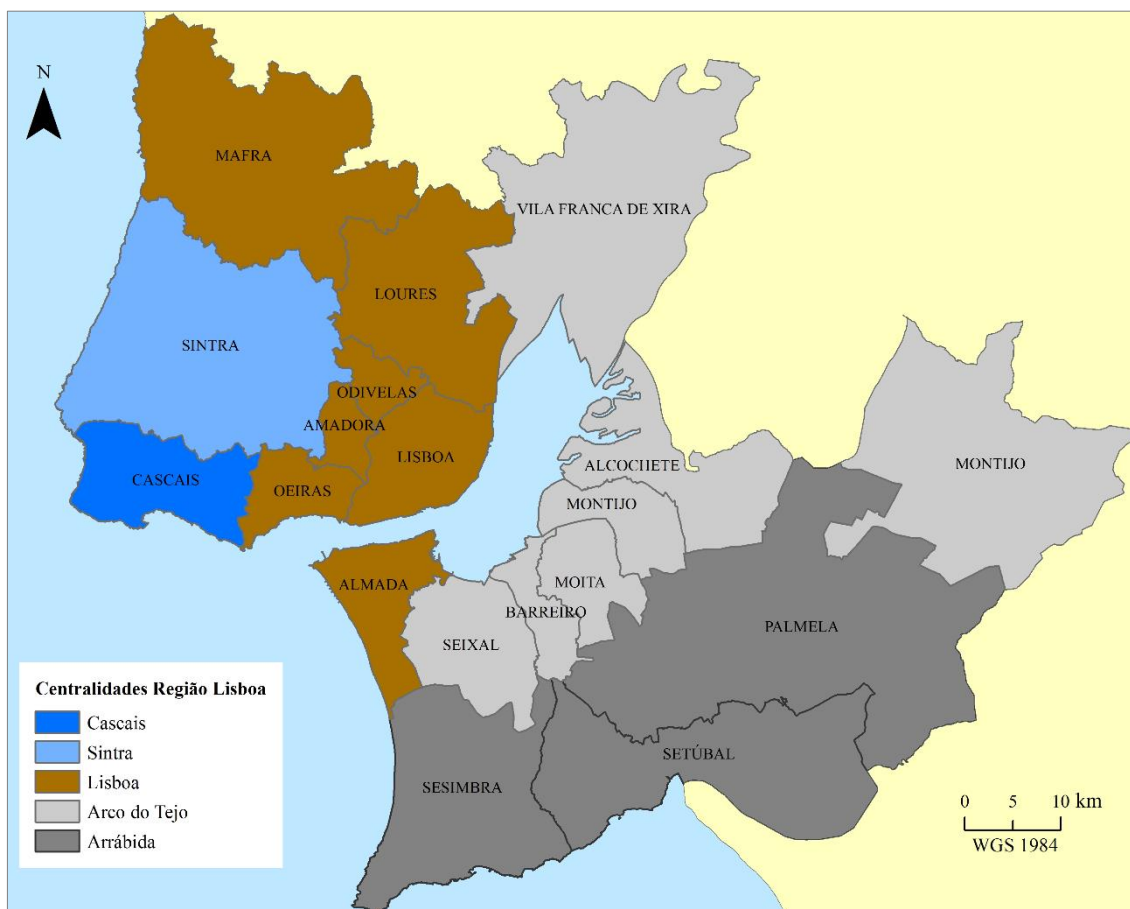
A construção de um modelo de regressão é um processo iterativo que envolve encontrar variáveis independentes capazes de explicar eficazmente a variável dependente que se pretende modelar (ou entender). Executando o modelo de regressão pode-se remover e/ou adicionar variáveis até encontrar o melhor modelo de regressão possível, determinando ao mesmo tempo quais as variáveis que são preditores eficazes.

Apesar de o processo de construção do modelo ser muitas vezes exploratório, nunca deve ser uma operação realizada de forma totalmente desinformada. Deve-se primeiramente identificar as variáveis explicativas candidatas, seja com o recurso a consultores, especialistas, ou conhecimento empírico. Devemos ser capazes de, antes da análise, indicar e justificar a relação esperada entre cada variável explicativa candidata e a variável dependente, e devem-se questionar os modelos onde essas relações não coincidem.

## 4. Métodos e Dados

### 4.1. Enquadramento geográfico: O turismo na cidade de Lisboa

Situada na Margem Norte da Área Metropolitana de Lisboa (AML), a cidade de Lisboa (Figura 4.1) tem uma área de aproximadamente 85 km<sup>2</sup> (CAOP 2012) e uma população residente de 547.773 indivíduos (INE, 2011).



**Figura 4.1: Centralidades da Região de Lisboa. Adaptado do PETRL (2015).**

Lisboa é considerada como a “...centralidade urbana mais importante da AML” (PROTAML, 2009, p.91)<sup>18</sup>. Segundo o Plano Estratégico para o Turismo na Região de Lisboa (PETRL) 2015 – 2019, a cidade é uma das centralidades turísticas da região, apresentando-se como “uma marca internacional forte que actua como âncora das restantes” (Cascais, Sintra, Arrábida e Arco do Tejo).

<sup>18</sup> No Diagnóstico Sectorial Sistema Urbano, Habitação e Equipamentos Constitucional refere-se à cidade de Lisboa como um sistema urbano de nível 1, salientando a “...concentração de equipamentos e serviços de nível superior e a centralidade conferida pelo sistema de transportes metropolitano”. (Plano Regional de Ordenamento do Território da Área Metropolitana de Lisboa 2009, p. 91).

De acordo com as Estatísticas do Turismo do Instituto Nacional de Estatística (INE, 2014), a AML possui 19,4% da capacidade de alojamento do país, destacando-se o município de Lisboa com 43.505 camas. Este valor não encontra paralelo no contexto da AML e só é superado a nível nacional pelo município de Albufeira que possui 47.685 camas.

Relativamente à procura turística, a AML concentra 24,95% do total das dormidas no país. A região continua a estar entre as primeiras escolhas dos principais mercados externos, sendo o principal destino para espanhóis, brasileiros, norte-americanos e italianos. No que respeita ao mercado interno, é o segundo destino de eleição dos residentes em Portugal (com 20,2% do total de dormidas). Lisboa é a cidade do país que apresenta o maior número de dormidas e de hóspedes, com mais de 9 e de 3,5 milhões, respectivamente (INE, 2014). A estada média dos hóspedes estrangeiros na cidade, é de 2,5 noites, abaixo da média da AML (2,6) e de Portugal Continental (3,1 noites).

Segundo o Inquérito Motivacional realizado em 2014 pelo Observatório de Turismo de Lisboa (OTL), tanto ao nível dos turistas estrangeiros como dos nacionais, a ‘*City & shortbreak*’ é considerada a maior motivação para visitar Lisboa.

A cidade de Lisboa destaca-se como um dos destinos favoritos de turismo urbano na Europa, tendo sido distinguida com vários prémios internacionais<sup>19</sup>. Lisboa patenteia uma oferta de actividades turísticas que se nutre das suas características intrínsecas, particularmente dos seus elementos culturais, históricos-patrimoniais e naturais, da sua valência comercial e de negócios, da sua diversidade de equipamentos (hotéis, restaurantes, centros de eventos, etc.) e da sua infra-estrutura de suporte (Brito-Henriques, 1996; PETRL, 2015).

Existem algumas áreas da cidade que participam com maior notoriedade na constituição da Lisboa Turística (Brito Henriques, 1996). Neste contexto, são apresentadas as principais áreas turísticas da cidade também denominadas como microcentralidades (MC), seguindo o zonamento proposto no PETRL 2015 – 2019 (Tabela 4.1 e Figura 4.2).

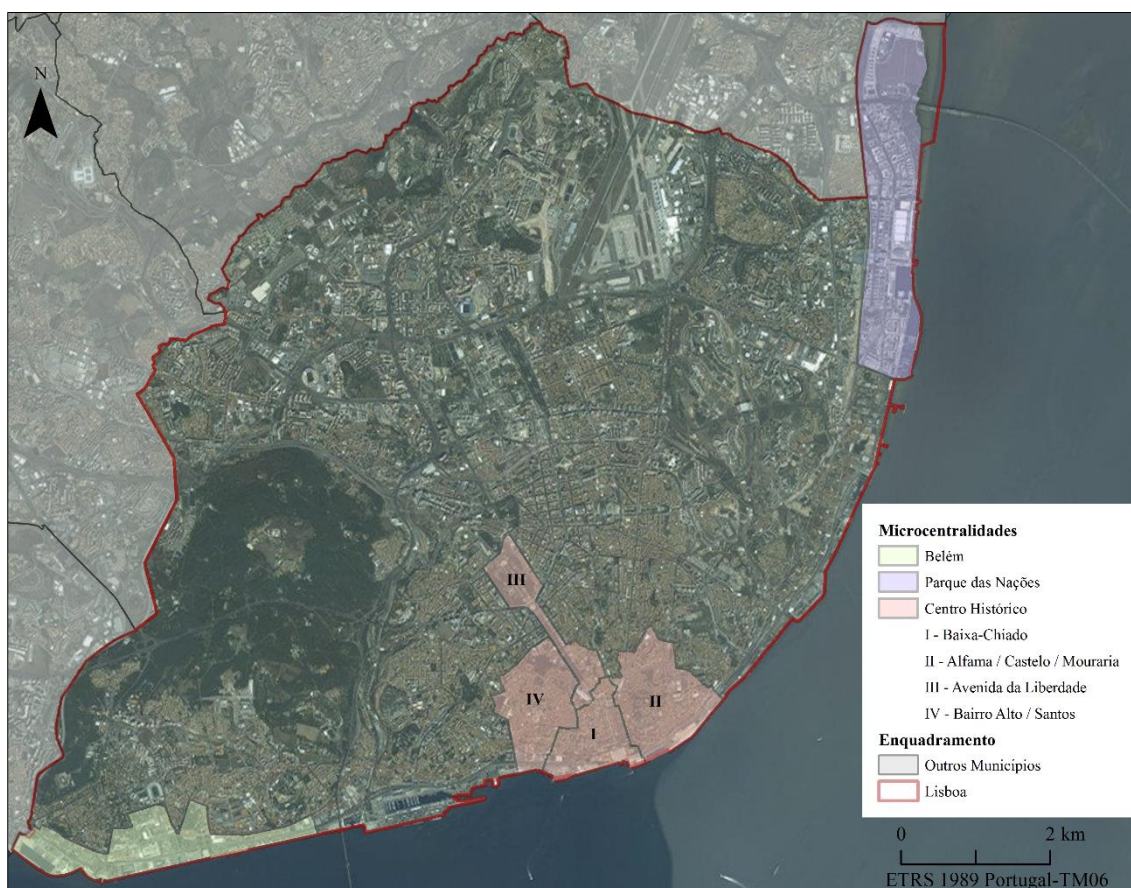
---

<sup>19</sup> *Europe's Leading City Break Destination 2013, 2010 e 2009; Europe's Leading Destination 2009; European Best Destination 2010*. Em *World Travel Awards*, consultado em Fevereiro, 2016. Url: <https://www.worldtravelawards.com/profile-8079-lisbon-portugal> e em *European Best Destinations*, consultado em Fevereiro, 2016. Url: <http://www.europeanbestdestinations.com/destinations>



**Tabela 4.1: Descrição das microcentralidades turísticas de Lisboa.**

Microcentralidade	Descrição	Relevância
Baixa-Chiado (BC)	Presença de monumentos e praças (Praças do Comercio, Restauradores, do Município) e de ruas pedonais dedicadas ao lazer (Rua Augusta). Concentra serviços de restauração, e locais de recreação (museu da Cerveja, do Chiado e <i>Lisboa Story Centre</i> ). Um dos bairros mais movimentados ao longo do dia, com animação de rua e programação cultural.	Histórico-patrimonial, cultural, religioso.
Alfama, Castelo e Mouraria (ACM)	Bairros mais antigos de Lisboa. Núcleo de festividades tradicionais (festas dos Santos populares) que expõem a Cultura Portuguesa. Distinguem-se os ícones da cidade (o Castelo de São Jorge e a Catedral do Sé), e os miradouros que oferecem panorâmicas do centro histórico da cidade e do rio.	
Bairro Alto, Cais do Sodré e Santos (BCS)	Área de excelência para o segmento jovem da cidade. Agrega uma diversidade de estabelecimentos (lojas de <i>design</i> e moda, galerias de arte, restaurantes, bares, etc.) que promovem a oferta de actividades associadas à vida nocturna, gastronomia, arte, cultura e moda.	Vida nocturna, gastronómica, cultural.
Marques de Pombal e Av. da Liberdade (PL)	Grande concentração de comércio de qualidade e de salas de espectáculos, com programação cultural diversificada. De salientar, os espaços verdes (parque Eduardo VII) e os eixos arborizados nos passeios que dão valor ambiental a identidade da cidade.	Comercial (luxo e sofisticação)
Parques das Nações (PN)	Espaço que atrai pela sua modernidade, ao nível de infra-estrutura, arquitectura e urbanismo. Pólo de atracção onde localizam-se o Oceanário, o Pavilhão do Conhecimento e outras infra-estruturas de animação (Teleférico, <i>shopping centre</i> , casino, teatro, etc.).	Modernidade
Belém (BE)	Importância arquitectónica e cultural. Localizam-se aqui, os atractivos mais visitados da cidade (Torre de Belém, Mosteiro dos Jerónimos) e outros equipamentos de referência como o Centro Cultural de Belém e o Museu dos Coches. Possui uma paisagem singular caracterizada pela combinação dos monumentos e dos espaços ajardinados e esplanadas localizadas junto ao rio.	Histórico-patrimonial, cultural, religioso.
Eixo Ribeirinho (ER)	Presença de espaços com funções de acolhimento de actividades náuticas turísticas e de recreio. Destaca-se o acesso pedonal à margem do rio e a fruição da paisagem ribeirinha.	Eixo de ligação entre as MC.



**Figura 4.2: Microcentralidades turísticas na cidade de Lisboa. Fonte: Plano Estratégico para o Turismo na Região de Lisboa 2015 – 2019.**

Segundo o inquérito às actividades dos turistas realizado em 2014 pelo OTL, para cerca de dois terços dos entrevistados a cidade de Lisboa foi o seu único destino de viagem (destes 64,1% são ‘primeiros visitantes’ [*First comers*]). A visita a monumentos, atracções ou museus, foi a actividade mais praticada pelos inquiridos, observando-se um maior ênfase no seio dos “estrepantes” (81,3%), face aos repetentes (*repeaters*) (66,3%).

Olhando para os resultados da Tabela 4.2, nota-se que existe uma predominância (em termos percentuais) do interesse pelos locais dentro da cidade, sendo os mais visitados, os situados nas MC: Belém e Baixa-Chiado. Verifica-se, também, a importância dos núcleos históricos-monumentais, Belém e Castelo, onde se localizam as atracções mais visitadas. De facto, as MC moldam o espaço vivido pelos turistas. A intensidade da actividade turística estende-se de Sul a Norte, na faixa ribeirinha entre Belém e a Praça de Comercio, passando pelo eixo entre a Baixa e a Av. Liberdade, e de forma pontual no Parque das Nações.

**Tabela 4.2: Locais de interesse e atracções visitadas, Inquérito às actividades dos turistas (OTL, 2014)<sup>20</sup>.**

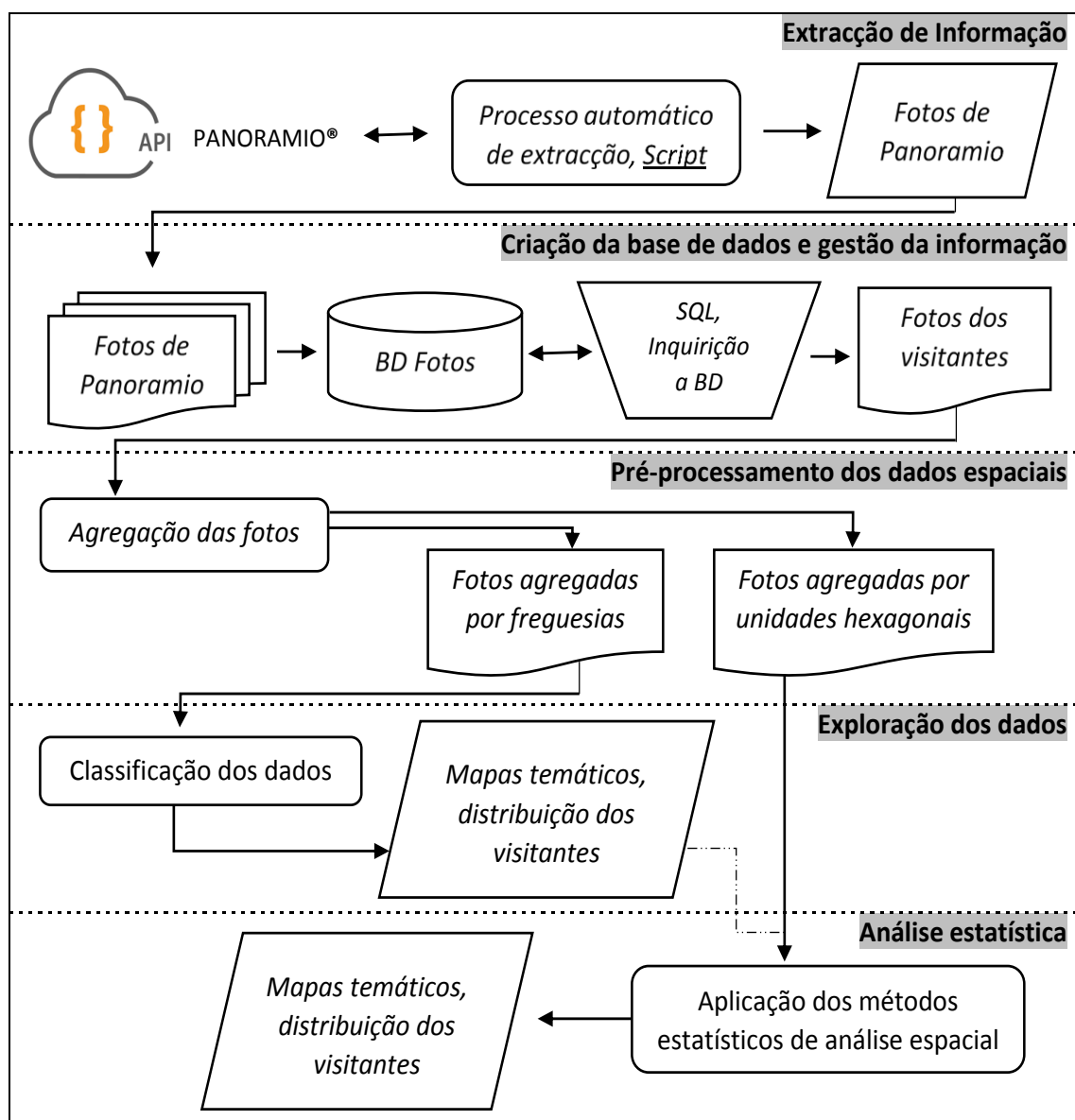
MC	Locais de interesse	%	Atracções visitadas	%	MC
BE	Belém	79,8	Torre de Belém	87,1	BE
BC	Baixa	74,9	Monastério dos Jerónimos	71,9	BE
PN	Parque das Nações	72,6	Castelo São Jorge	63,5	ACM
BC	Terreiro do Paço	72,1	Padrão dos Descobrimentos	63,1	BE
PL	Avenida Liberdade	70,3	Oceanário	40,6	PN
BCS	Bairro Alto	67,2	Sé de Lisboa	37,1	ACM
ACM	Alfama	64,8	Palácio Ajuda	28,8	BE
BC	Chiado	57,0	Pavilhão do Conhecimento	23,8	PN
-	Sintra	30,8	Centro Cultural de Belém	21,6	BE
-	Cascais	24,2	Museu Gulbenkian	14,7	PL
-	Estoril	24,1	Casino Lisboa	11,8	PN
-	Fátima	16,1	Museu de Arte Antiga	10,5	BCS
-	Arrábida	8,7	Museu dos Coches	10,4	BE
-	Setúbal	7,4	Museu do Chiado	8,6	BC
-	Sesimbra	6,2	Museu do Design (MUDE)	2,8	BC
-	Ericeira	3,3	Lisboa Story Centre	2,2	BC

#### 4.2. Procedimento metodológico para a análise espacial

Como já referido na secção 3.3. o modelo conceptual de análise espacial inclui o desenvolvimento das seguintes cinco etapas: i) Recolha de informação; ii) Pré-processamento dos dados; iii) Exploração dos dados; iv) Aplicação dos métodos estatísticos; e iv) Interpretação dos resultados (O'Sullivan e Unwin, 2010; Charlton, 2008).

Nesta fase, a estrutura do projecto é apresentada e explicada de forma esquemática (Figura 4.3), sendo orientada numa lógica sequencial de etapas processuais desde a aquisição da informação até a etapa da aplicação dos métodos estatísticos.

<sup>20</sup> O inquérito foi realizado a turistas estrangeiros dentro e fora da cidade de Lisboa. A informação da tabela corresponde aos resultados das entrevistas em Lisboa (mais do 80% do total de entrevistas). Não são apresentados as atracções mais visitadas situadas fora da cidade. A tabela original é anexada (Anexo 1).



**Figura 4.3: Mapa cognitivo do processo de análise espacial utilizado no estudo.**

Em seguida, apresenta-se uma síntese das etapas apontadas no esquema (Figura 4.3):

- **Extração da informação:** Os dados procedem: i) da Carta Administrativa Oficial de Portugal (CAOP 2015) e ii) da informação associada às fotos *geotagged* disponíveis na rede de partilha de fotos denominada “Panoramio”. A CAOP foi descarregada do repositório oficial da Direcção Geral do Território (DGT) de Portugal. A informação correspondente às fotos *geotagged* foi extraída através da API da rede de partilha. A API proporciona o acesso às fotos (e aos seus atributos)

carregadas pelos utilizadores. O pedido de informação foi realizado através de um *http request*<sup>21</sup> dirigido ao servidor do Panoramio.

Para descarregar a informação de toda área que compreende a cidade de Lisboa, tornou-se necessário automatizar o processo de “pedido de informação”. Para o efeito, operacionalizaram-se os pedidos através de um *script* concebido em linguagem Python<sup>22</sup>. A informação foi recebida em vários ficheiros de texto em formato *.json*<sup>23</sup> que foram posteriormente convertidos em formato *.xls*<sup>24</sup>.

- Criação da base de dados e gestão da informação: A informação foi estruturada em tabelas e importada para uma base de dados (BD). A BD facilitou o tratamento da informação para dois fins: i) Eliminação de redundância de informação; ii) Organização e preparação da informação antes da sua migração para ambiente SIG. O primeiro ponto envolve a validação da informação extraída da rede e, o segundo, a segmentação do conjunto de dados para a análise.

Do conjunto total de fotos foi necessário diferenciar aquelas tiradas pelos visitantes e pelas pessoas locais. Para o efeito, identificou-se o número de dias que existiam entre a primeira e a última foto carregada pelos utilizadores dentro da área de estudo. Foram considerados como visitantes, os utilizadores que tinham carregado fotos durante um período inferior a três dias (procurou-se que este critério fosse ao encontro do perfil-tipo de visita turística à cidade, predominantemente *short-break*, *i.e.*, 2,5 noites) e, como locais, os utilizadores que tinham carregado fotos durante um período mais lato. O tratamento da informação e a segmentação dos dados foram concretizados dentro um Sistema de Gestão de Base de Dados (SGBD) que proporciona o ambiente necessário para a manipulação e consulta dos dados.

- Pré-processamento dos dados espaciais: Esta etapa é desenvolvida já em ambiente SIG e refere-se especialmente à agregação dos dados (fotos dos visitantes) quer

---

<sup>21</sup> A comunicação entre um servidor e um “cliente” ocorre, através de um processo de pedido/resposta. O “cliente” envia uma mensagem de pedido indicando o que é preciso, o servidor procura a informação e envia a resposta (Heilmann e Francis, 2007; Mateu, 2010).

<sup>22</sup> Python é uma linguagem de programação, sob o modelo *open-sourced*, criada por Guido Van Rossum em 1990. Em Python *help*, consultado em Abril, 2016. url: <https://docs.python.org/2/license.html>

<sup>23</sup> JSON (*JavaScript Object Notation*) é um formato para armazenamento e transferência de informação. Em w3schools *tutorials*, consultado em Abril, 2016. url: <http://www.w3schools.com/json/>

<sup>24</sup> XLS refere-se ao formato dos ficheiros de Microsoft Excel.

para análise exploratória, quer para análise estatística. Para a etapa de exploração, os dados foram agregados ao nível da freguesia<sup>25</sup>, com base na divisão administrativa da CAOP 2015. Para a análise estatística, tal como realizado por Palomares *et al.* (2015), a agregação foi efetuada a partir de uma malha regular de hexágonos com unidades (células) de  $\approx 2.346\text{m}^2$ , que corresponde à dimensão média das subsecções (*i.e.* quarteirões urbanos) na cidade de Lisboa. O valor atribuído a cada uma das unidades hexagonais corresponde à soma das observações individuais dentro de cada célula. Ainda a respeito do pré-processamento dos dados para a análise estatística, também se realizou uma segmentação temporal do conjunto de fotografias para os anos 2007, 2009 e 2012, com a respectiva agregação espacial por unidades hexagonais.

Aquando da migração da informação para ambiente SIG, todos os dados foram organizados dentro de uma base de dados espacial ou *Geodatabase*<sup>26</sup> (.gdb).

- Exploração dos dados: A exploração dos dados envolveu as técnicas em ambiente SIG referidas na secção 3.3.1. para a visualização da distribuição espacial e a avaliação da não estacionariedade (espacial) dos dados. Aqui, é apresentado o conjunto (*dataset*) das fotografias dos visitantes (não agregadas), e dois mapas temáticos criados a partir da agregação das fotos pelas áreas das freguesias de Lisboa. São também descritas, através de infografia, algumas características dos dados no que diz respeito à proporção das fotografias e dos utilizadores (visitantes) por mês e por ano.
- Análise estatística: Nesta parte, foram aplicados os métodos estatísticos citados na secção 3.3.2., conforme os objetivos do trabalho, por um lado, para a análise da distribuição espacial dos visitantes e a identificação dos locais típicos e atípicos de concentração (considerando não só a relação espacial entre as observações mas também a componente temporal dos dados) e, por outro lado, para explorar as relações entre o padrão observado e outras variáveis explicativas (24 variáveis seleccionadas).

---

<sup>25</sup> Classificação de mais baixa hierarquia dentro dos níveis possíveis no quadro de divisão administrativa estabelecida de acordo com a Constituição Portuguesa segundo o catálogo de entidades da CAOP (Instituto Geográfico Português, 2012).

<sup>26</sup> É a designação para uma base de dados espacial segundo o programa hospedeiro, ArcGIS. Trata-se de um repositório de informação geográfica que suporta diferentes tipos de objetos (Arctur e Zeiler, 2004).

A apresentação e interpretação dos resultados são tópicos desenvolvidos no capítulo 6.

#### 4.2.1. Informação base para o estudo

A metodologia de análise espacial preconizada neste trabalho foi operacionalizada em ambiente SIG, segundo uma estrutura de dados vectorial, com o sistema de projecção de coordenadas PT-TM06/ETRS89. Os dados de base utilizados para o desenvolvimento deste trabalho são descritos na Tabela 4.3.

**Tabela 4.3: Dados de base para a análise espacial.**

Variável	Geometria	Descrição	Fonte / Sistema de referência (origem)
CAOP (Carta administrativa Oficial de Portugal)	Polígonos	A carta que delimita e demarca as circunscrições administrativas do País, do ano 2015.	Direcção - Geral do Território, Portugal <sup>27</sup> . PT-TM06/ETRS89 <sup>28</sup>
<i>Geotagged</i> fotos localizadas na cidade de Lisboa	Pontos	Fotos georreferenciadas partilhadas pelos utilizadores da rede “Panoramio”, entre os anos 2006 e 2014.	Rede social de partilha de fotos “Panoramio” <sup>29</sup> . WGS84 <sup>30</sup> .

#### 4.2.2. Programas e aplicações empregados

Ao longo do trabalho, utilizaram-se vários programas e aplicações em cada uma das etapas, os quais permitiram a execução das diferentes tarefas. O *software* SIG utilizado para manusear a informação geográfica foi o ArcGIS 10.3., sendo particularmente aproveitados os módulos *Mapping Clusters* e *Analyzing Patterns* da caixa de ferramenta *Spatial Statistics Tools*. A Tabela 4.4 faz a síntese dos programas e aplicações afectos ao estudo.

<sup>27</sup> Consultado em Fevereiro, 2016. url: [http://www.dgterritorio.pt/cartografia\\_e\\_geodesia/cartografia/](http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/)

<sup>28</sup> A Associação Internacional de Geodesia (IAG), através da sua subcomissão *European Reference Frame* (EUREF), recomenda a utilização do sistema global de referência a ser utilizado no âmbito da União Europeia seja coincidente com o *International Terrestrial Reference System* (ITRS) de 1989. Este sistema deverá ser fixado à parte estável da Placa Euro-Asiática, sendo designado por Sistema de Referência Terrestre Europeu 1989 (*European Terrestrial Reference System* 1989 - **ETRS89**). A agência *EuroGeographics* recomenda a utilização da projecção cartográfica Transversa de Mercator (TM) para escalas superiores a 1:500 000. Em 2006 (06) foram definidos os parâmetros desta projecção para Portugal (PT). Este é um sistema de coordenadas cartesianas expressas em metros.

<sup>29</sup> Consultado em Dezembro, 2015. url: <http://www.panoramio.com/api/widget/api.html>

<sup>30</sup> O *World Geodetic System* de 1984 (WGS 84) é um sistema de coordenadas geográficas normalmente expressas em graus. A diferença entre o WGS 84 e o ITRS 89 é de poucos centímetros.

**Tabela 4.4: Lista dos programas utilizados no trabalho.**

Etapa	Programa	Descrição/Utilidade	Fonte
1	Python 2.7.11.	O programa que suporta a linguagem de programação Python.	<a href="https://www.python.org/downloads/">https://www.python.org/downloads/</a>
	WingIDE 5.1.	Editor que reúne as ferramentas de apoio para o desenvolvimento de programas Python.	<a href="http://wingware.com/downloads/wingide-101">http://wingware.com/downloads/wingide-101</a>
2	Opal-Converter JSON	Aplicação para converter ficheiros. json para .xls; .csv	<a href="http://www.skytopia.com/software/opal-csv-converter/json-to-csv.htm">http://www.skytopia.com/software/opal-csv-converter/json-to-csv.htm</a>
	Access 2013	Sistema de gestão de base de dados para a manipulação e consulta de dados, desenvolvido pela Microsoft.	Licença comercial
3	Create Hexagon Tessellation	Aplicação que cria uma malha regular de hexágonos sobre uma área, compatível com ArcGIS.	<a href="http://www.arcgis.com/home/item.html?id=03388990d3274160afe240ac54763e57">http://www.arcgis.com/home/item.html?id=03388990d3274160afe240ac54763e57</a>
4 – 5	ArcGIS Pro	<i>Software</i> da ESRI que proporciona ferramentas adicionais de análise espacial ( <i>Space time cube, Emerging hot spot analysis</i> ).	Licença comercial
Todas	ArcGIS 10.3.	<i>Software</i> da ESRI <sup>31</sup> . SIG para trabalhar com informação geográfica.	Licença comercial

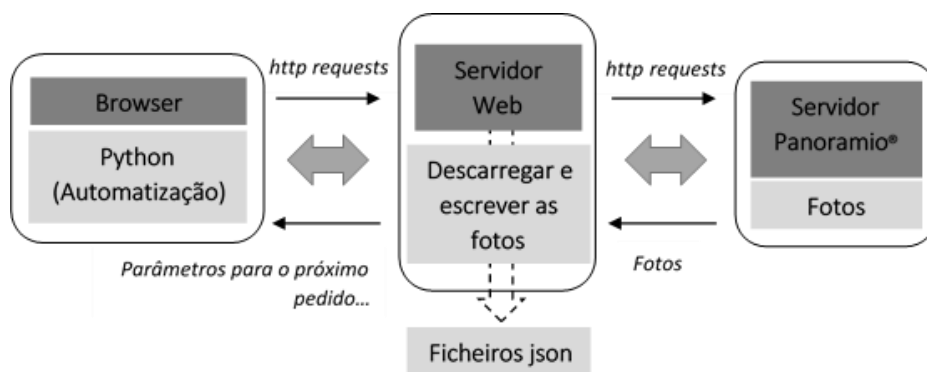
### 4.3. Extracção dos dados

#### 4.3.1. O processo de requerimento de informação

Como já referido na secção 4.3.1., a API da rede de partilha de fotos “Panoramio” proporciona o acesso à base de dados das fotografias com o intuito de serem utilizadas em aplicações web. Para além das fotos (as imagens), tem-se acesso à informação que lhes está associada. De acordo com o protocolo estabelecido pela API, o processo de extracção de informação deve ser encaminhado através de um pedido (*http request*) onde se especificam os detalhes relativos aos dados necessários e se indica o caminho da conexão ou de intercâmbio de dados. O servidor do Panoramio recebe o pedido, realiza a pesquisa e, em resposta, envia a informação solicitada em formato JSON (Figura 4.4).

<sup>31</sup> ESRI (*Environmental Systems Research Institute, Inc.*) é a empresa que desenvolve os productos ArcGIS.





**Figura 4.4: Processo para extrair informação do servidor da Panoramio.**  
Adaptado de Kádár e Gede (2013).

No caso do Panoramio, o pedido é expresso como<sup>32</sup>:

**[http://www.panoramio.com/map/get\\_panoramas.php?set=public&from=0&to=20&minx=-180&miny=-90&maxx=180&maxy=90&size=medium&mapfilter=true](http://www.panoramio.com/map/get_panoramas.php?set=public&from=0&to=20&minx=-180&miny=-90&maxx=180&maxy=90&size=medium&mapfilter=true)**

A primeira parte da “linha de pedido” (a negrito) identifica a localização da informação consoante a estrutura do directório do servidor e contém encriptado o método *http*<sup>33</sup> “get” que sugere a requisição da informação. No resto da “linha de pedido” declaram-se os parâmetros que determinam qual a informação a solicitar. Cada um dos parâmetros está separado pelo símbolo “&” (AND) e pode ser alterado seguindo as normas da API. Os parâmetros e as suas opções de configuração são descritos na Tabela 4.5.

A API do Panoramio estabelece um limite para a extracção de dados, sendo possível retirar informação até um máximo de 500 fotos por cada pedido. No caso de uma área ter mais fotos disponíveis, devem-se realizar outros pedidos que continuem com a extracção da informação, declarando a sequência das fotos nos parâmetros “from” e “to”.

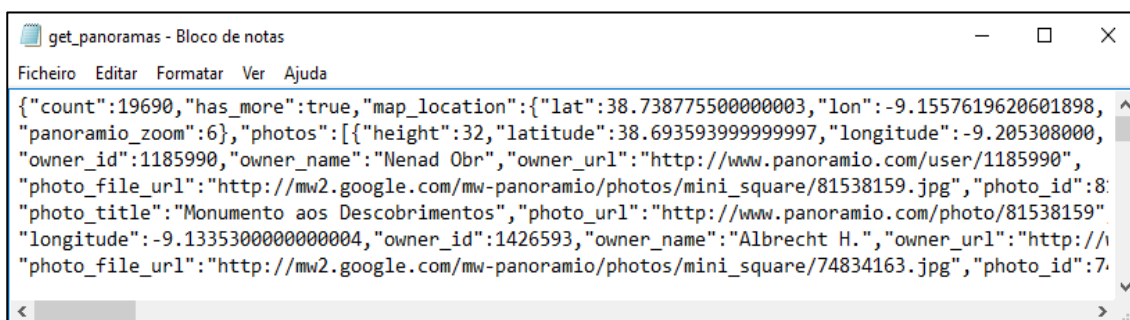
Depois de realizado o pedido, o servidor envia a resposta em formato JSON que pode ser descarregado para o computador. Na Figura 4.5 é apresentado um exemplo de conteúdo da resposta (neste caso, o ficheiro é visualizado através do editor de texto do Microsoft Windows). A resposta possui um cabeçalho que contém detalhes sobre a área especificada no requerimento e um corpo que inclui a informação relativa às fotos.

<sup>32</sup> Em Panoramio *help*, consultado em Dezembro, 2015. url: <http://www.panoramio.com/api/data/api.html>

<sup>33</sup> *Hypertext Transfer Protocol* é um protocolo de transferência de dados utilizado na comunicação entre os *browsers* e os servidores (Mateu, 2010).

**Tabela 4.5: Parâmetros do pedido de informação.**

Parâmetro	Valores	Descrição
Set	Public Full ID do utilizador	Procura somente as fotos populares. Procura todas as fotos. Procura fotos pertencentes a um utilizador em particular.
From (X) – To (Y)	Valores numéricos	Define o número de fotos a solicitar (Y - X). O valor “0” representa a foto mais recente carregada na rede. Por exemplo, ao declarar-se from=0 to=20, extraem-se as últimas 20 fotos carregadas na rede e ao declarar-se from=20 to=40 extraem-se as 20 fotos que precedem o primeiro conjunto, e assim sucessivamente.
Minx Miny Maxx Maxy	Coordenadas geográficas em graus decimais. (WGS84).	Delimitam a área ( <i>bounding box</i> <sup>34</sup> ) de onde se pretende extrair as fotografias. Corresponde às coordenadas dos cantos inferior esquerdo e superior direito (Mínimos longitude e latitude, Máximos longitude e latitude, respectivamente).
Size	Original Medium/Small Thumbnail Square Mini_square	Diz respeito às diferentes dimensões em que podem ser apresentadas as fotos. Refere-se exclusivamente à configuração das fotos no momento da resposta. Estes valores não alteram a informação relativa à dimensão original das fotos.
Mapfilter	True - False	É um filtro aplicado à foto, quando apresentada sobre um mapa, para melhorar o seu aspecto.



**Figura 4.5: Visualização da resposta do servidor Panoramio num editor de texto.**

No cabeçalho encontram-se quatro indicadores:

- **Count:** Revela o número total de fotos dentro da área indicada no pedido mas não representa a quantidade de fotos disponíveis para descarregar;
- **Has\_more:** Informa se existem, ou não, mais fotos disponíveis para descarregar. Os valores possíveis são “True” e “False”;
- **Map\_location:** Mostra as coordenadas do centroide da área delimitada. São coordenadas geográficas expressas em graus decimais no sistema WGS84;
- **Panoramio\_zoom:** Este indicador sugere o nível de *zoom* onde estão alojadas as fotos para visualização no sítio *web* da rede. Na interface de exploração do

<sup>34</sup> O *bounding box* (*bbox*) refere-se a uma área definida por coordenadas latitude e longitude.

Panoramio, as fotos com maior popularidade<sup>35</sup> são visualizadas em níveis de *zoom* menores e as fotos menos representativas vão aparecendo com o aumento do *zoom*.

A informação das fotos é apresentada a seguir ao cabeçalho. Toda a informação referente às fotos encontram-se dentro do separador “Photos” e cada registo de foto está delimitado por “{ }”. Para cada foto é gerado um identificador único que acompanha os restantes atributos. São nove os atributos associados a cada foto:

- O identificador único da foto;
- O título da foto. É a descrição textual da foto inserida pelo utilizador no momento do *upload*. O conteúdo do título pode também incluir palavras-chaves (*hashtags*);
- O *Uniform Resource Locator* (URL) da foto. O endereço na rede onde se pode encontrar a foto;
- As coordenadas geográficas no sistema de referência WGS84, identificadas pelos separadores: “latitude” e “longitude”;
- As dimensões da foto original, em largura e comprimento;
- A data de *upload*;
- O identificador único do proprietário da foto (o utilizador que carrega a foto);
- O nome do proprietário conforme a informação do seu perfil;
- O URL do perfil do utilizador.

Com o intuito de perceber o processo de requisição de informação, foram realizados alguns testes de pedidos através do navegador web Google Chrome. Estas experiências foram operacionalizadas de forma manual e permitiram concluir que:

- Nos ficheiros de resposta, o valor do indicador “count” refere-se ao total de fotos localizadas na área e que, na maioria dos casos, este não é necessariamente igual à quantidade de fotos disponíveis para descarregar.
- Para determinar que todas as fotos disponíveis numa área foram descarregadas, deve-se ter especial atenção ao resultado do indicador “has\_more”. Logo ao realizarem-se alguns pedidos na mesma área, dando continuidade aos valores dos

---

<sup>35</sup> O índice de popularidade de uma foto é determinado por uma combinação de mais de 20 variáveis que incluem: o número de vezes em que a foto foi visualizada (na plataforma de Panoramio), o número de comentários recebidos, o número de vezes que outros utilizadores marcaram a foto como favorita, a resolução da foto em pixéis, etc. No entanto, o Panoramio não divulga todas as variáveis consideradas no cálculo para evitar tentativas de manipulação do índice. Em Panoramio *help*, consultado em Dezembro, 2016. url: [http://www.panoramio.com/help/understanding\\_popularity](http://www.panoramio.com/help/understanding_popularity)

parâmetros “from” e “to” (como explicado na Tabela 4.5), verifica-se que o estado deste indicador muda para “false” quando já não existe informação para retirar.

- O volume de fotos disponíveis para descarregar é significativo. Assim, para conseguir descarregar as fotos de toda a área que compreende a cidade de Lisboa, o processo de requisição necessitava de ser automatizado.

#### **4.3.2. A automatização do requerimento de informação**

Para a automatização era necessário criar um processo que fosse capaz: i) de enviar os pedidos de forma sequencial mas que modificasse continuamente os parâmetros do *htt request*, particularmente, aqueles que indicam qual as fotos a extrair (“from” e “to”); e ii) de suspender a comunicação iterativa “pedido/resposta” quando não exista mais informação disponível para retirar.

Da experiência realizada por Kádár e Gede (2013), determinou-se que o processo podia ser desenvolvido através de uma aplicação ou “programa” que operacionalizasse os requerimentos. Deve-se referir que os autores, no seu trabalho, descrevem o processo de extracção de forma geral sem especificar os procedimentos necessários para a automatização dos pedidos.

Procurou-se então, em repositórios/bibliotecas digitais, por *scripts*<sup>36</sup> ou programas de código aberto relacionados com a extracção de dados de redes sociais. No repositório GitHub<sup>37</sup>, encontrou-se um *script* desenhado para extrair informação da rede Panoramio, cuja estrutura podia ser modificada em função dos interesses deste trabalho. O *script* original<sup>38</sup> foi desenvolvido em linguagem Python e executava as seguintes tarefas:

- A extracção de fotos (exclusivamente das imagens) existentes dentro uma área. A configuração inicial do código permitia definir a localização de apenas uma área de cada vez.

---

<sup>36</sup> Um *script* ou código é uma lista de comandos ou sequência de instruções que podem ser interpretadas por um programa (Maruch e Maruch, 2006; Hetland, 2008).

<sup>37</sup> O GitHub é uma plataforma web de colaboração e partilha de códigos para programadores. Em GitHub *guides*, consultado em Dezembro, 2015. url: <https://guides.github.com/>

<sup>38</sup> O *script* original é apresentado no Anexo 2. Foi extraído do repositório GitHub em Dezembro, 2016. Em: [https://raw.githubusercontent.com/automatingosint/osint\\_public/master/image\\_sites/auto\\_image\\_search\\_extended\\_1.py](https://raw.githubusercontent.com/automatingosint/osint_public/master/image_sites/auto_image_search_extended_1.py)

- A criação de uma página web em linguagem HTML<sup>39</sup> para a visualização das fotos extraídas.

O *script* foi analisado para perceber a sua estrutura, a sequência das instruções e o funcionamento dos seus comandos. Neste código fizeram-se algumas alterações simples e acrescentaram-se duas instruções que modificaram a sua aplicação original.

Nesta parte, é apresentada e explicada a estrutura do *script*, sendo identificadas cada umas das modificações realizadas para ajustar o seu funcionamento (realçadas a laranja). Para a edição do código foi necessário a instalação do Python 2.7.11. e de um programa de edição, o WingIDE 5.1.

Esta parte é destinada à apresentação do código:

---

```

1. import requests
2. import json
3. import os
4.
5. corners = ["38.706264,-9.13776,38.709015,-9.135611",
6.           "38.708942,-9.147431,38.711693,-9.145282"]
7.
8. for x in corners:
9.     print(x)
10.
11.     search_name = "lisboa"+x
12.
13.     # we use the coordinates
14.     y1,x1,y2,x2 = x.split(",")
15.
16.     #
17.     # Function responsible for sending requests to Panoramio API
18.     #
19.     def send_panoramio_request(start,end):
20.
21.         api_url = "http://www.panoramio.com/map/get_panoramas.php?set=full&"
22.         api_url += "from=%d&to=%d&" % (start,end)
23.         api_url += "minx=%s&miny=%s&maxx=%s&maxy=%s&" % (x1,y1,x2,y2)
24.         api_url += "size=medium&mapfilter=false"
25.
26.         response = requests.get(api_url)
27.
28.         if response.status_code == 200:
29.
30.             # convert to a dictionary
31.             search_results = json.loads(response.content)
32.
33.             return search_results
34.
35.         # there was a problem
36.         return None
37.
38.     #
39.     # Use the Panoramio API to get all pictures

```

<sup>39</sup> É uma linguagem de programação utilizada para a criação de páginas web (Mateu, 2010).

```

40.     #
41.     def get_all_panoramio_pictures():
42.
43.         photo_list = []
44.         search_start = 0
45.
46.         # send the initial request
47.         search_results = send_panoramio_request(search_start,search_start+499)
48.
49.         # if there was an error return nothing
50.         if search_results is None:
51.             print "[*] Error retrieving photos."
52.             return
53.
54.         # add the current results to our list
55.         photo_list.extend(search_results['photos'])
56.
57.         print "[*] Retrieved %d photos..." % (len(photo_list))
58.
59.         # while there are more photos to retrieve
60.         while search_results['has_more'] is True:
61.
62.             # we increase the search count by 499
63.             search_start += 499
64.
65.             search_results = send_panoramio_request(search_start,search_start+499)
66.
67.             if search_results is not None:
68.                 photo_list.extend(search_results['photos'])
69.
70.                 print "[*] Retrieved %d photos..." % (len(photo_list))
71.
72.         # return all of our photos
73.         return photo_list
74.
75.     #
76.     # Write out the list of photos
77.     #
78.     def write_photo_list(photo_list):
79.
80.         if not os.path.exists("%s" % search_name):
81.             os.mkdir("%s" % search_name)
82.
83.         fd = open("%s/%s.html" % (search_name,search_name),"wb")
84.         fd.write("<html><head></head><body>")
85.
86.         # walk through the list of photos and add them to our log file
87.         for photo in photo_list:
88.             fd.write("<a target='_blank' href='https://maps.google.com/?q=%f,%f'><img src='%s' title='%s' alt='%s' border='0'></a><br/>\r\n" %
89.                 (photo['latitude'],photo['longitude'],photo['photo_file_url'],photo['photo_title'],photo['photo_id']))
90.
91.         # close the html file
92.         fd.write("</body></html>")
93.         fd.close()
94.
95.         # added code for thesis
96.         with open("%s.json" % search_name,"wb") as fd:
97.             fd.write(json.dumps(photo_list))
98.
99.             print "[*] Wrote out %s.json for Luis!" % search_name
100.
101.         return
102.
103.         # grab the full list of photos from panoramio

```

```
104.         photo_list = get_all_panoramio_pictures()
105.
106.         write_photo_list(photo_list)
```

---

Nas primeiras linhas (1 - 3) faz-se referência aos módulos<sup>40</sup> *requests*, *json* e *os* para ter acesso as suas funcionalidades e incorporá-las no programa. Os módulos estão disponíveis na biblioteca de Python, *pypi*<sup>41</sup>, e devem ser instalados antes da sua aplicação. As funções destes módulos são:

- O módulo *requests* administra a comunicação “pedido/resposta” e encaminha os pedidos de forma automática através do navegador web.
- O módulo *json* é utilizado para decodificar a resposta do servidor do Panoramio (em formato JSON) e interpretar o seu conteúdo.
- O módulo *os* disponibiliza as funções do sistema operativo (neste caso, Windows). Vai permitir que o programa possa executar tarefas como: criar, abrir e editar ficheiros ou navegar através de um directório, etc.

Conforme já referido, o programa original estava desenhado para apenas extrair informação de uma área. No caso de haver diferentes áreas de interesse, o programa devia ser executado varias vezes (introduzindo as coordenadas de cada uma das áreas, uma de cada vez). Face a esta limitação, no contexto deste trabalho foram acrescentadas funções que permitem gerir a extracção de informação de diferentes áreas. Assim, na linha 5 é definida uma variável denominada “corner” onde são listadas as coordenadas das áreas de interesse. A função expressa na linha 8 indexa cada grupo de coordenadas referente às áreas distintas e repete iterativamente o processo de extracção de informação para todas as áreas enumeradas.

Entre as linhas 19 e 24 define-se o pedido. Nos parâmetros “from” e “to” faz-se uma ligação a outras linhas do código (44 e 47) que indica quais as fotos a retirar, sendo neste exemplo as 499 fotos mais recentes da área, segundo a indexação do Panoramio (cf. Tabela 4.5). Os parâmetros que se debruçam sobre a localização da área (minx, miny, maxx, maxy) estão ligados à variável “corner” que possui a listagem de coordenadas das áreas de interesse.

---

<sup>40</sup> Um módulo é um programa desenhado para executar tarefas específicas. As suas funcionalidades podem ser utilizáveis em outros programas (neste caso, dão suporte ao código para extrair informação de Panoramio). Em *Python tutorials*, consultado em Dezembro, 2016. url: <https://docs.python.org/3/tutorial/modules.html>

<sup>41</sup> *Pypi* é o repositório *open-source* de módulos de Python. Consultado em Dezembro, 2015. url: <https://pypi.python.org/pypi/py>

Desde a linha 28 até à 36, estabelece-se o envio do pedido conforme os parâmetros definidos nas linhas anteriores. O pedido é enviado e verifica-se o seu estado (28), *i.e.*, se foi aceite o não. A ser positivo, a resposta é recebida e decodificada (31). A partir deste ponto, o programa consegue interpretar o conteúdo da resposta do servidor do Panoramio.

A instrução para retirar toda a informação disponível numa determinada área é declarada entre as linhas 41 e 73. Logo após receber a primeira resposta, o programa procura o resultado do indicador “has\_more”. Caso este seja “false”, a informação relativa às fotos identificadas no separador “photos” é importada para um ficheiro de texto, e termina o processo de extracção para essa área. Se o valor do parâmetro for “true” (linha 60), a informação da primeira resposta é importada para um ficheiro de texto mas o processo continua sobre a mesma área. Neste caso, é enviado um novo pedido (65) alterando, somente, os valores dos parâmetros “from” e “to”. Para um pedido seguinte, o valor do “from” é definido do seguinte modo (linha 63): o valor de procura inicial mais 499 ( $0 + 499$ ), e o valor do “to” (linha 47) é igual ao valor de procura “from” mais 499 ( $499 + 499$ ). A nova resposta é decodificada e interpretada pelo programa e a informação relativa às fotos deste pedido é incluída no mesmo ficheiro de texto. Este processo é iterativo até a resposta do estado do indicador “has\_more” ser “false”.

A parte do código entre as linhas 78 e 93 é dedicada a criação da página web para a visualização das fotos (as imagens) descarregadas. Como já referido, esta instrução cumpre os objectivos originais do programa.

A segunda das instruções adicionadas neste trabalho é apresentada entre as linhas 96 e 99. Esta função é a que permite guardar a informação descarregada num novo ficheiro. O programa cria um ficheiro de texto com extensão *.json* designado como “Lisboa + as coordenadas da área do processo em curso”. Neste ficheiro, é importada toda a informação das fotos (o identificador único, as coordenadas, o título, etc. - Ver secção 4.3.1.). No final do processo, o ficheiro aparece disponível no directório de trabalho. Para cada área de interesse, o programa cria um ficheiro com a respectiva informação.

Para sua utilização, o código deve ser transposto para um ficheiro de texto e gravado, utilizando a extensão “.py” (para que possa ser identificado pelo Python). Recomenda-se que o ficheiro do programa seja guardado num directório (pasta) independente, devido a que a informação a extrair vai ser encaminhada automaticamente para esta localização.



#### **4.3.3. *Extracção dos dados para o estudo***

A área de estudo foi segmentada em áreas de menor dimensão e, dentro dessas subáreas, foram executados os pedidos para extrair informação. Para o primeiro teste de extracção, utilizando já o processo automático, segmentou-se a área de estudo em subáreas de 1000m x 1000m. A informação extraída de cada uma das subáreas foi agrupada, tendo como resultado um total de 49.048 registos. De forma a verificar se foi retirada toda a informação disponível, efectuou-se um outro teste, neste caso, segmentando a área de estudo em subáreas de 500m x 500m. De forma semelhante ao primeiro teste, a informação extraída foi agrupada, tendo como resultado 55.229 registos.

Ao identificar e apagar os registos duplicados em cada grupo de dados, o total de registos dos testes resultou em 43.052 para o primeiro e em 48.456 para o segundo. A diferença no número de fotos entre os dois testes, evidenciou que o volume total de imagens extraídas pode variar conforme a dimensão das subáreas.

Com o intuito de obter a maior quantidade de fotos possível, efectuaram-se mais testes (em total vinte e dois), mudando a dimensão das subáreas<sup>42</sup>. Logo, realizou-se uma comparação entre os vinte dois grupos de dados extraídos, para determinar quais as fotos repetidas e únicas em cada um deles. O conjunto de dados utilizados para este trabalho, resultou da junção das fotos únicas resultantes de cada um dos testes. A comparação entre os diferentes grupos de dados e o conjunto final de dados é apresentado na secção 4.4.1.

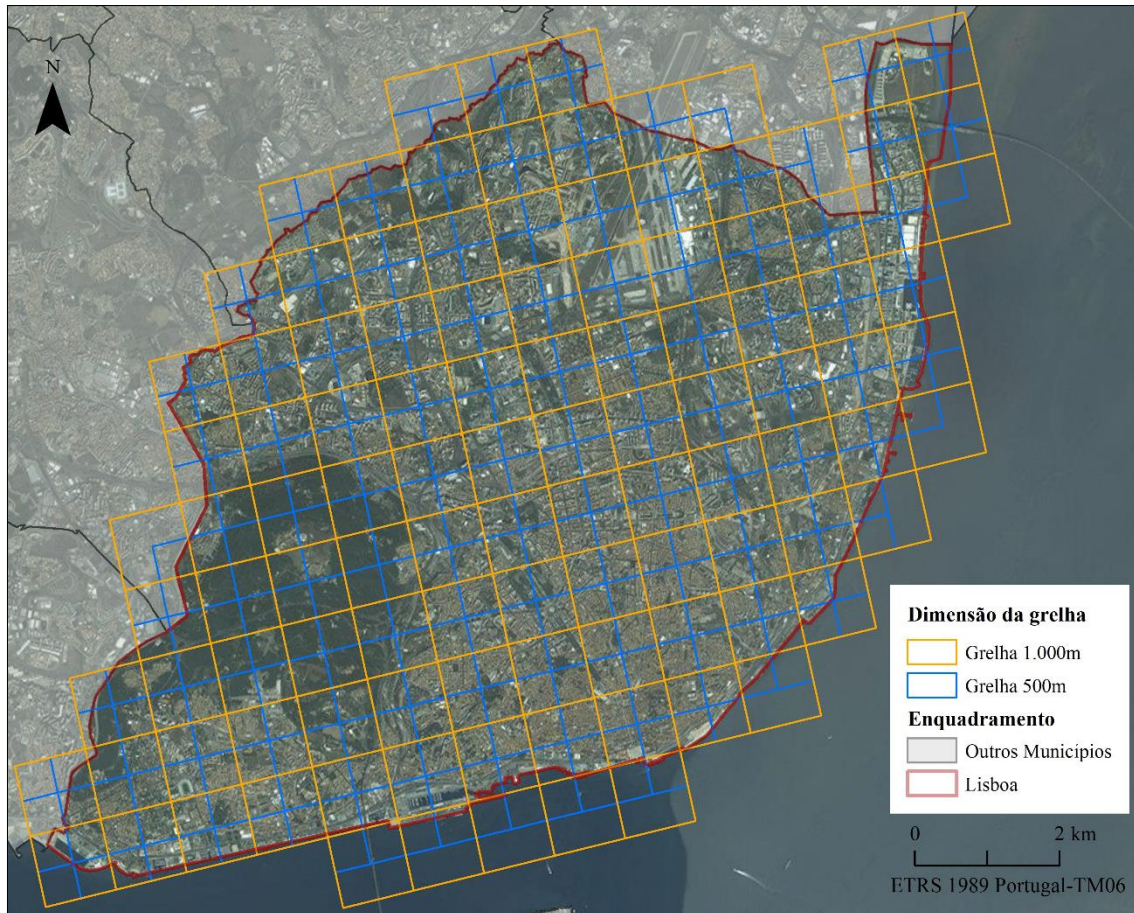
O processo para extrair informação inicia-se com a determinação das coordenadas das subáreas. Em seguida, essas coordenadas são organizadas e inseridas na variável “corner” (no código), e o código é executado. O procedimento para determinar e organizar as coordenadas das subáreas, segundo o formato estabelecido no código é apresentado na secção seguinte. Os detalhes sobre a execução do código são apresentados na secção 4.3.3.2.

---

<sup>42</sup> A quantidade de fotos extraídas em cada teste e o número de registos duplicados são apresentados na Tabela 4.6.

#### 4.3.3.1. Determinação das coordenadas das subáreas

Sobre a área que compreende a cidade de Lisboa foram delimitadas diferentes malhas regulares, cada uma com a sua respectiva dimensão de célula (Figura 4.6).



**Figura 4.6: Malhas regulares.**

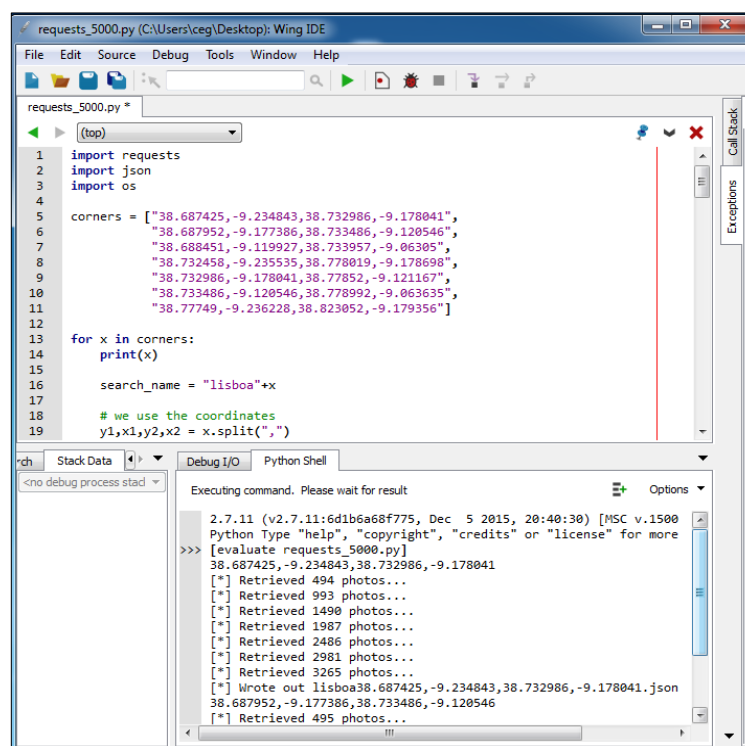
Os vértices das células foram tomados como referência para determinar as coordenadas das subáreas. Este processo foi realizado em ambiente SIG, utilizando as seguintes ferramentas: *Create fishnet*, *Vertex to points*, *Extract coordinates*. A primeira cria a malha regular (nos parâmetros define-se a dimensão de célula pretendida). A segunda identifica os vértices de cada polígono (células) e cria pontos sobre eles e, a última, extrai as coordenadas desses pontos.

Como resultado, obtém-se um novo *layer* (ou camada de informação), cuja tabela de atributos possui as coordenadas dos pontos nos vértices das células (cada ponto tem um código que o associa à respectiva célula). Nesta fase, trabalhou-se com coordenadas geográficas decimais no sistema de referência WGS84, uma vez que a API do Panoramio só reconhece este tipo de coordenadas. A informação da tabela de atributos foi exportada

para um ficheiro *.csv*<sup>43</sup>. O conteúdo do ficheiro foi importado para uma folha de cálculo de Excel, para continuar com o processo. De seguida, identificaram-se, para cada polígono, as coordenadas dos cantos: inferior esquerdo e superior direito. Assim, as coordenadas foram organizadas segundo o formato estabelecido no *script* (minx, miny, maxx, maxy) e ordenadas de forma a ter uma lista. Como se observa na Figura 4.7, o conjunto de coordenadas de uma subárea é enumerado entre aspas (“ ”) e cada conjunto é separado por vírgulas (,).

#### 4.3.3.2. Executando o código para a extracção dos dados

Como já foi referido, as coordenadas das subáreas devem ser inseridas na variável “corner”. Para isso, o ficheiro que contém o código é aberto através do editor (WingIDE 5.1.), e a lista de coordenadas já organizadas é colocada dentro dos parênteses rectos da variável. Antes de executar o código, o ficheiro deve ser guardado.



**Figura 4.7: Interface do programa editor WingIDE 5.1.**

É possível executar directamente o código a partir da interface do programa editor, quando o código está a ser executado pode-se acompanhar o processo de extracção na janela inferior denominada “Python shell” (Figura 4.7). Nesta janela, aparecem os avisos

<sup>43</sup> *Comma-separated values* ou ficheiro de texto com valores separados por vírgula.

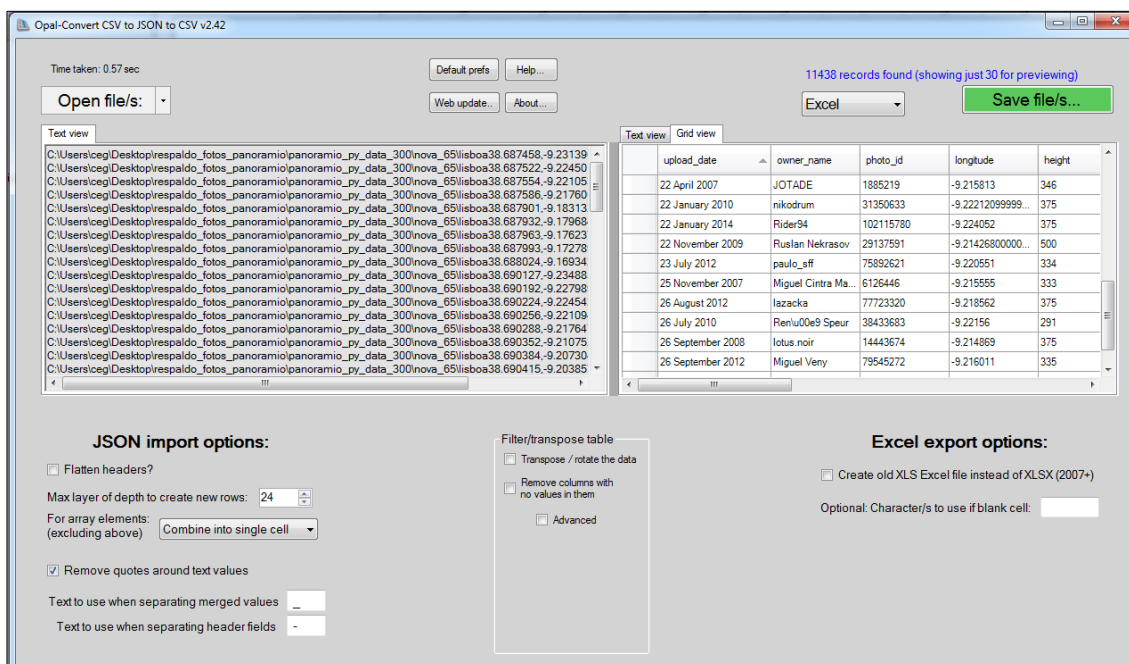
que indicam qual a subárea onde a extracção está a ser realizada, o número de fotos a descarregar e a conclusão do processo de extracção na subárea.

Depois de se extrair os dados de uma subárea, os resultados são escritos num novo ficheiro de texto (com extensão .json) e o processo de extracção continua na próxima subárea, seguindo a lista de coordenadas. Os ficheiros com a informação das fotos das subáreas aparecem disponibilizados no directório estabelecido (o mesmo directório onde se localiza o código fonte).

O código também pode ser executado através de uma consola de Python. Neste caso, simplesmente deve-se dar “double click” sobre o ficheiro que contém o código. Aparecerá automaticamente a consola de Python, onde se pode acompanhar o processo com os mesmos avisos antes mencionados. Os novos ficheiros que contêm a informação vão ficando disponíveis conforme o processo de extracção decorre. Quando termina a extracção de todas as subáreas, a consola de Python fecha-se automaticamente.

#### ***4.3.3.3. Conversão dos ficheiros de resposta (texto) para folhas de cálculo***

Os ficheiros de resposta com a informação das fotos referentes a cada uma das subáreas, em formato .json, foram convertidos para folhas de cálculo (Excel neste caso), através da aplicação “Opal-converter” (Figura 4.8). A aplicação permite converter vários ficheiros simultaneamente e integrá-los numa única folha de cálculo (formato .xls). Para efectuar a conversão, deve-se escolher a opção “Open files from folder” da interface da aplicação e identificar a pasta onde estão hospedados os ficheiros a converter.



**Figura 4.8: Interface da aplicação “Opal Converter”, para converter os ficheiros de JSON para XLS.**

#### 4.4. Criação da base de dados e gestão da informação

A informação retirada da rede de partilha foi armazenada numa base de dados para facilitar a sua organização (considerando o volume) e garantir a integridade dos dados. A base de dados é do tipo relacional.

A opção pelo modelo de dados relacional é importante porque os seus conceitos básicos são acessíveis e gerais, e porque a sua delineação não depende de nenhum programa informático específico. O conjunto de dados foram manipulados usando um Sistema de Gestão de Base de Dados (SGBD), tendo-se optado neste caso pelo Access 2015.

A informação resultante dos testes de extracção (num total vinte e dois) que estava anteriormente organizada em folhas de cálculo, foi depois importada para a nova base de dados, mantendo a mesma estrutura. Na Figura 4.9, é apresentada a organização das tabelas na base de dados, discriminadas segundo o teste de extracção.

As vinte e duas tabelas importadas possuíam, cada uma, sete campos (colunas) que continham a informação de alguns dos atributos das fotos referidos na secção 4.3.1. Cada registo (ou tuplo) refere-se à informação de uma foto (Figura 4.10).



**Figura 4.9: Tabelas importadas para a base de dados.**

Id_foto	Id_utilizador	latitude	longitude	titulo_foto	url_foto	data_upload
10399084	481208	38.707490	-9.173230	Trenes de Philadelphia importados por PORTUGAL.	www.panoramio.com/photo/10399084	5/19/2008
10399092	481208	38.698975	-9.179163	Vista parcial de Lisboa	www.panoramio.com/photo/10399092	5/19/2008
10401314	1745789	38.709985	-9.142030	S. Carlos	www.panoramio.com/photo/10401314	5/19/2008
10401329	1745789	38.710713	-9.141955	Brasileira summer night	www.panoramio.com/photo/10401329	5/19/2008
104065321	4941785	38.713497	-9.140604	Lisboa; Rossio; estatua de Dom Pedro IV	w.panoramio.com/photo/104065321	3/8/2014
104065321	4941785	38.713497	-9.140604	Lisboa; Rossio; estatua de Dom Pedro IV	w.panoramio.com/photo/104065321	3/8/2014
1040831	214509	38.711115	-9.130358	In The Streets of Lisboa - Alfama	www.panoramio.com/photo/1040831	2/26/2007
104132568	7227151	38.742595	-9.145410	A vision through the raindrops on a glass ceiling	w.panoramio.com/photo/104132568	3/9/2014
104132785	7227151	38.744327	-9.150667	Ritmos da noite	w.panoramio.com/photo/104132785	3/9/2014

**Figura 4.10: Modelo de tabela importada para a base de dados.**

#### 4.4.1. Determinação do conjunto de dados a utilizar no estudo

##### 4.4.1.1. Comparação da informação extraída nos distintos testes

Posteriormente à importação das folhas de cálculo, já em ambiente do SGBD, procurou-se: i) Eliminar os registros duplicados em cada uma das tabelas e ii) Comparar cada uma das tabelas com todas as outras. A comparação de informação foi realizada para identificar e agrupar os registros únicos extraídos nos diferentes testes, e obter a maior quantidade possível de fotos da área de estudo.

Em cada tabela, utilizando uma expressão em linguagem SQL (4.1), seleccionaram-se todos os registros únicos<sup>44</sup>, os quais foram utilizados para criar novas tabelas.

<sup>44</sup> Na Tabela 4.6, é apresentado o número de registros identificados como duplicados, em cada teste.

```

SELECT distinct(Id_foto),*
into tabela_nova
from tabela_original

```

(4.1)

É de referir que nas novas tabelas, o campo que contém o identificador da foto (ID\_foto) foi declarado como “chave primária”. A “chave primária” distingue cada registro de todos os outros e permite estabelecer as ligações/relações entre tabelas. Estas tabelas mantinham a mesma estrutura da tabela apresentada na Figura 4.10 mas, em cada uma, foi acrescentado um novo campo denominado “grelha” para identificar a dimensão da célula do respectivo teste de extracção. Para criar o campo (4.2) e inserir a informação nas suas células (4.3), utilizaram-se as expressões:

```

Alter table tabela_nova
add column grelha text

```

(4.2)

```

Update tabela_nova Set grelha = '5000'
Where grid Is Null

```

(4.3)

Uma vez eliminados os registros duplicados, efectuou-se a comparação de informação entre tabelas, de forma sequencial desde o primeiro até o último teste. A comparação foi operacionalizada com instruções expressas em linguagem SQL. De modo a exemplificar o processo de comparação de informação, apresenta-se a comparação entre as fotos extraídas no primeiro e segundo teste. Tendo como referencia os códigos de cada foto, identificaram-se e apagaram-se as fotos do conjunto de dados “B” (fotos extraídas do teste, cuja malha regular possuía células de 4.500m<sup>2</sup>) que apareciam repetidas no conjunto de dados “A” (fotos extraídas do teste, cuja malha regular possuía células de 5.000m<sup>2</sup>). Para isso, aplicou-se a expressão (4.4).

```

delete from tabela_b
where exist(
select Id_foto from tabela_a
where tabela_a.Id_foto=tabela_b.Id_foto
)

```

(4.4)

A seguir, ambas as tabelas “A” e “B” (sem os registros repetidos) foram unificadas (*join*) numa nova tabela, utilizando a expressão (4.5). Os registros desta nova tabela “AB” foram confrontados com a informação obtida no seguinte teste (as fotos extraídas no teste cuja malha regular possuía células de 4.000m<sup>2</sup>), e assim sucessivamente até chegar aos dados do último teste.



```

SELECT bd.*
into tabela_ab
From (Select
Id_foto, Id_utilizador, latitude, longitude,
titulo_foto, url_foto, data_upload, grelha
FROM tabela_a
UNION Select
Id_foto, Id_utilizador, latitude, longitude,
titulo_foto, url_foto, data_upload, grelha
FROM tabela_b
) bd

```

(4.5)

Das 13.949 fotos extraídas no primeiro teste, após a comparação e agrupamento das novas fotos, ficou-se com um total de 75.112 fotos (Tabela 4.6). Relativamente aos dados obtidos em cada teste, evidenciou-se que o número de fotos extraídas foi aumentando à medida que a dimensão da célula da malha regular foi diminuindo, o que pode estar relacionado com a forma em que as fotos estão organizadas para visualização na plataforma do Panoramio:

*When you explore the world by using Panoramio's map or Google Earth, you want to see the best photos of a place first! That's why the most popular photos are visible at higher zoom levels. You'll see the other photos as you zoom into a location, with the least popular photos appearing at the lowest zoom levels (Panoramio help<sup>45</sup>).*

Contudo, se considerarmos a diferença entre o número de fotos extraídas no teste 22 e o total de fotos, pode-se concluir que existem fotos hospedadas em níveis inferiores de *zoom* que não estão disponíveis para extrair em níveis superiores. Portanto, para se obter a maior quantidade possível de fotos de uma determinada área dever-se-á replicar o processo adoptado neste estudo.

Esta experiência pode guiar futuras tentativas de extracção de informação, uma vez que a API do Panoramio não menciona nenhum detalhe sobre a influência da dimensão da célula, ou melhor, da área, no processo de extracção de fotos. A síntese do número de fotos extraídas, duplicadas, únicas, e novas para cada teste e o acumulado total, é apresentada na Tabela 4.6.

---

<sup>45</sup> Consultado em Dezembro, 2015. url: [http://www.panoramio.com/help/understanding\\_popularity](http://www.panoramio.com/help/understanding_popularity)



**Tabela 4.6: Síntese do número de fotos extraídas por teste.**

Teste	Dimensão célula (m)	Fotos					
		Extraídas	Duplicadas	Únicas	Novas	Acumuladas	%
1	5.000	14.144	195	13.949	0	13.949	18,6
2	4.500	14.181	145	14.036	2.826	16.775	3,8
3	4.000	17.471	138	17.333	3.267	20.042	4,3
4	3.500	20.992	204	20.788	3.688	23.730	4,9
5	3.000	24.694	335	24.359	3.762	27.492	5,0
6	2.800	24.098	538	23.560	1.953	29.445	2,6
7	2.600	28.274	402	27.872	2.726	32.171	3,6
8	2.400	29.982	612	29.370	2.529	34.700	3,4
9	2.200	36.341	716	35.625	6.270	40.970	8,3
10	2.000	38.014	518	37.496	4.322	45.292	5,8
11	1.800	38.407	762	37.645	1.614	46.906	2,1
12	1.600	38.579	790	37.789	1.846	48.752	2,5
13	1.400	41.566	809	40.757	1.574	50.326	2,1
14	1.200	48.902	967	47.935	2.867	53.193	3,8
15	1.000	49.048	5.996	43.052	4.535	57.728	6,0
16	900	52.095	978	51.117	2.047	59.775	2,7
17	800	55.208	757	54.451	2.076	61.851	2,8
18	700	57.253	705	56.548	2.101	63.952	2,8
19	600	58.400	791	57.609	1.464	65.416	1,9
20	500	55.229	6.773	48.456	1.014	66.430	1,3
21	400	63.824	1.043	62.781	2.628	69.058	3,5
22	300	66.151	8.627	57.524	6.054	<b>75.112</b>	8,1
							100,0

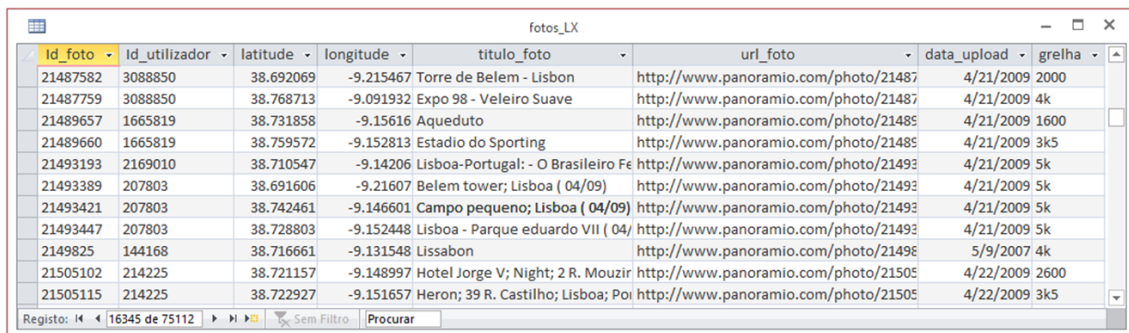
#### 4.4.1.2. Diferenciação das fotos tiradas por visitantes e locais

O método para diferenciar as fotos tiradas por visitantes e locais assenta nos trabalhos realizados por Girardin *et al.* (2008), Kádár (2014) e García-Palomares *et al.* (2015). Os autores calcularam a diferença de dias entre a marca temporal (*timestamps*) da primeira e última foto carregadas pelos utilizadores, dentro da área de estudo. Se o número de dias entre a primeira e última foto excedesse a média do tempo de visita do destino, as fotos eram consideradas como pertencentes a locais, ao invés, se o número de dias estava dentro do período proposto, então eram consideradas como pertencentes a visitantes.

Como referido na secção 2.3.2.1., as datas de referência para a diferenciação, constituem as datas de carregamento de fotos à rede de partilha que, nalguns casos, podem

não coincidir com as datas em que as fotos foram geradas pelos utilizadores. No caso da rede Panoramio, apenas se tem acesso à data de *upload*, não tendo nenhuma informação (disponível para carregar) sobre a data em que foi capturada a foto.

A partir da tabela base que contém a informação das 75.112 fotos (Figura 4.11), gerou-se outro conjunto de tabelas para sintetizar a informação e, assim, conseguir a diferenciação das fotos. À semelhança das operações para comparar a informação entre tabelas da secção anterior, as instruções para diferenciar os dados foram expressas em linguagem SQL.



Id_foto	Id_utilizador	latitude	longitude	titulo_foto	url_foto	data_upload	grelha
21487582	3088850	38.692069	-9.215467	Torre de Belem - Lisbon	http://www.panoramio.com/photo/21487	4/21/2009 2000	
21487759	3088850	38.768713	-9.091932	Expo 98 - Veleiro Suave	http://www.panoramio.com/photo/21487	4/21/2009 4k	
21489657	1665819	38.731858	-9.15616	Aqueduto	http://www.panoramio.com/photo/21485	4/21/2009 1600	
21489660	1665819	38.759572	-9.152813	Estadio do Sporting	http://www.panoramio.com/photo/21485	4/21/2009 3k5	
21493193	2169010	38.710547	-9.14206	Lisboa-Portugal: - O Brasileiro Fe	http://www.panoramio.com/photo/21493	4/21/2009 5k	
21493389	207803	38.691606	-9.21607	Belem tower; Lisboa ( 04/09)	http://www.panoramio.com/photo/21493	4/21/2009 5k	
21493421	207803	38.742461	-9.146601	Campo pequeno; Lisboa ( 04/09)	http://www.panoramio.com/photo/21493	4/21/2009 5k	
21493447	207803	38.728803	-9.152448	Lisboa - Parque eduardo VII ( 04/	http://www.panoramio.com/photo/21493	4/21/2009 5k	
2149825	144168	38.716661	-9.131548	Lissabon	http://www.panoramio.com/photo/21498	5/9/2007 4k	
21505102	214225	38.721157	-9.148997	Hotel Jorge V; Night; 2 R. Mouzir	http://www.panoramio.com/photo/21505	4/22/2009 2600	
21505115	214225	38.722927	-9.151657	Heron; 39 R. Castilho; Lisboa; Poi	http://www.panoramio.com/photo/21505	4/22/2009 3k5	

**Figura 4.11: Tabela base de fotos da área de estudo.**

Tendo como referência a data de *upload* e o código identificador dos utilizadores, criou-se uma nova tabela, organizando as fotos por utilizador e por ano, e determinando a quantidade de dias entre a primeira e ultima foto carregada na rede. Utilizando a expressão (4.6), identificou-se para cada utilizador, as datas mínima e máxima das suas fotos. Em seguida, calculou-se a diferença de dias e a quantidade de fotos.

```
SELECT fotos_LX.Id_utilizador, Min([data_upload]) as min_data,
Max([data_upload]) as max_data,
DateDiff('d',Min([data_upload]),Max([data_upload]),1,1) AS n_dias,
Year([data_upload]) AS ano, Count(fotos_LX.data_upload) AS n_fotos (4.6)
into utilizadores_diffdata
From fotos_LX
Group by fotos_LX.Id_utilizador, Year([data_upload])
```

A tabela resultante, designada de “utilizadores\_diffdata”, ficou armazenada na base de dados, para ser utilizada no decorrer do processo de diferenciação. No conjunto de dados, encontraram-se um total de 7.116 utilizadores (considerando locais e visitantes). Analisando a informação, identificaram-se utilizadores com fotos carregadas em anos diferentes, e que o intervalo (de dias) entre as fotos carregadas em cada ano

variava significativamente. Não podendo diferenciar se estes utilizadores são locais ou, de facto, visitantes recorrentes, estas ocorrências foram descartadas da base de dados utilizando a expressão (4.7):

```
Select id_utilizador, count_ano
into utilizadores_1vez
from (
Select distinct(id_utilizador), count(ano) as count_ano
from utilizadores_diffdata
group by id_utilizador
)
where count_ano = 1
```

(4.7)

Sem a informação dos utilizadores com fotos carregadas em mais do que um ano, a quantidade total de registos diminuiu para 6.152. Aplicando a expressão (4.8), seleccionaram-se unicamente os utilizadores que apresentavam valores inferiores a 3 no campo “n\_dias” (número de dias entre a primeira e a ultima foto carregada na rede), o que corresponde à estada média dos visitantes de Lisboa (secção 4.1.). Com estes registos criou-se uma nova tabela apresentada na Figura 4.12, denominada “utilizadores\_1vinf3”.

```
SELECT id_utilizador, min_data, max_data, n_dias, ano, n_fotos
into utilizadores_1vinf3
from utilizadores_1vez
where n_dias < 3
```

(4.8)

A tabela<sup>46</sup> manteve a mesma estrutura da tabela “utilizadores\_diffdata” mas só continha os registos dos utilizadores considerados como visitantes, num total de 5.082 utilizadores.

---

<sup>46</sup> Para representar o verdadeiro número de dias, somou-se uma unidade a todos os valores do campo “n\_dias” da tabela. O valor de “0” representaria um dia, “1” seria igual a dois dias, e assim sucessivamente.

Id_utilizador	min_data	max_data	n_dias	n_fotos	ano
1271220	4/26/2009	4/26/2009	0	1	2009
1271441	5/5/2011	5/6/2011	1	3	2011
1272161	1/2/2009	1/2/2009	0	1	2009
127316	12/11/2007	12/11/2007	0	1	2007
1274441	1/14/2008	1/16/2008	2	5	2008
127447	5/7/2011	5/7/2011	0	2	2011
12750	1/29/2007	1/29/2007	0	3	2007
127532	1/27/2007	1/27/2007	0	1	2007
1276400	1/12/2008	1/12/2008	0	2	2008

**Figura 4.12: Diferencia de datas das fotos carregadas pelos visitantes.**

Tendo como referencia o código identificador dos utilizadores, efectuou-se uma ligação entre esta tabela “utilizadores\_1vinf3” e a tabela base que continha a informação das fotos (apresentada na Figura 4.11). Utilizando a expressão (4.9), seleccionaram-se todas as fotos da tabela base que pertenciam aos utilizadores considerados como visitantes.

```
SELECT fotos_LX.*
into fotos_utilizadores1v_inf3
from fotos_LX, utilizadores_1vinf3
where fotos_LX.Id_utilizador = utilizadores_1vinf3.Id_utilizador
```

(4.9)

Com as fotos seleccionadas criou-se uma nova tabela denominada “fotos\_utilizadores1v\_inf3” que corresponde ao conjunto de dados utilizados para a análise. A tabela resultante continha 17.604 fotos tiradas por utilizadores considerados como visitantes, cada uma com os seus atributos, como apresentado na Figura 4.11. Mais detalhes sobre este conjunto de dados são apresentados na secção 5.2.

#### **4.4.2. Migração do conjunto de dados para uma BD espacial, em ambiente SIG**

A tabela contendo a informação das fotos dos visitantes, hospedada na base de dados em Access, foi exportada para uma folha de Excel e guardada num novo ficheiro (.xls). Este tipo de ficheiros, à imagem dos provenientes do Access, podem ser interpretados pelo programa ArcGIS. Neste caso, foi necessário recorrer a este passo intermédio devido às dificuldades apresentadas pelo ArcGIS em interpretar alguns dos campos das tabelas em Access.

Já num ambiente SIG, na interface da aplicação *desktop* do ArcMap, gerou-se um novo projecto e a folha de cálculo com os dados das fotos foi importada para o espaço de

conteúdos (*Table of Contents* [TOC]) do projecto. A folha de cálculo aparece como uma tabela de atributos na TOC mas sem nenhuma representação gráfica no “visualizador” do ArcMap. Para disponibilizar os dados no visualizador, recorreu-se à ferramenta denominada “Display X Y data”. Neste processo deve-se indicar: i) quais os campos da tabela de atributos que possuem a informação das coordenadas “Y” e “X” dos pontos (os campos denominados “Latitude” e “Longitude”, respectivamente); e ii) o sistema de referência das coordenadas, sendo coordenadas geográficas WGS 84. Deste modo, os dados aparecem projectados no visualizador, tendo associada a tabela de atributos inicialmente importada.

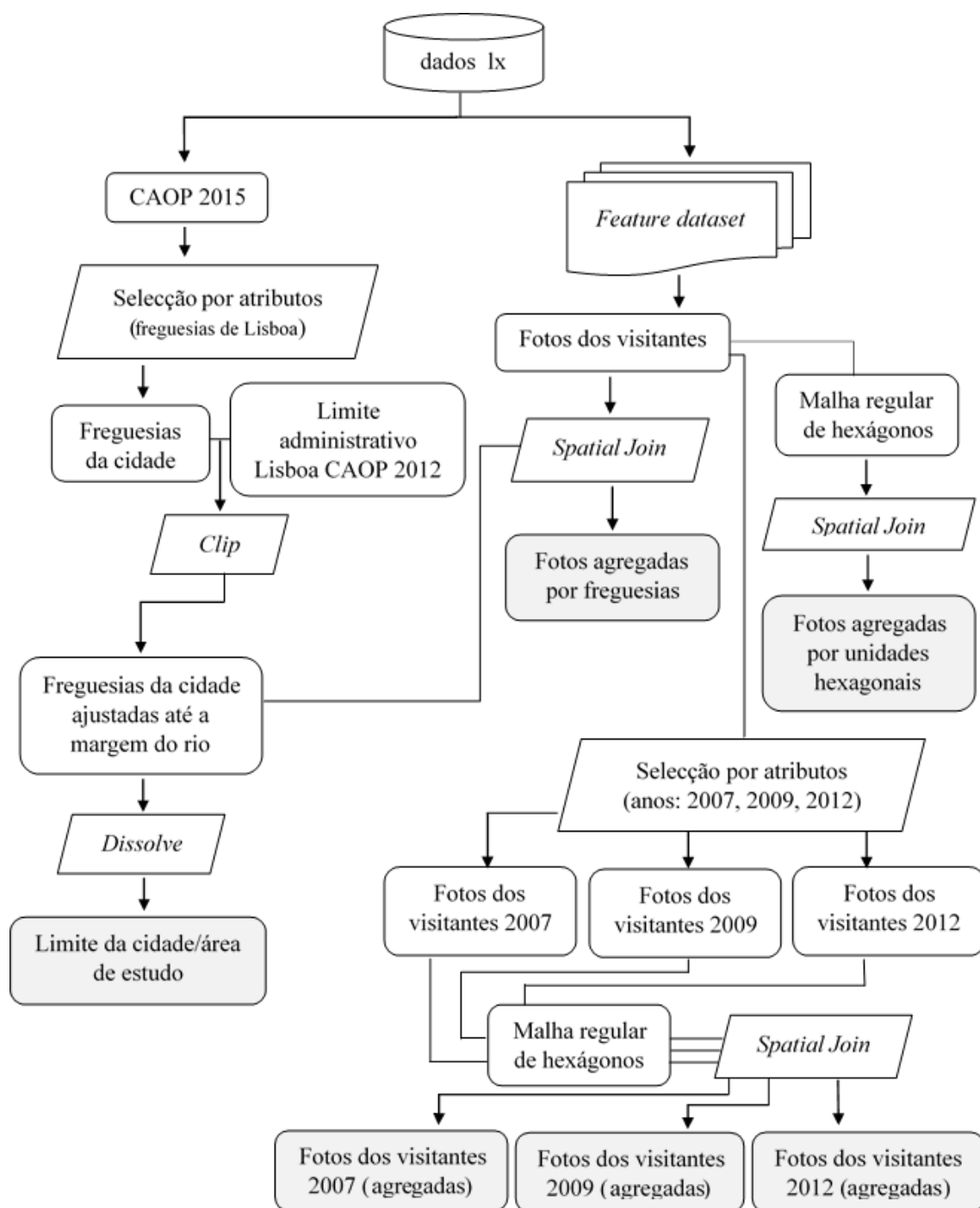
Posteriormente, criou-se uma *geodatabase* e uma *feature dataset*<sup>47</sup>, para armazenar todos os dados a utilizar no estudo, utilizando as ferramentas *Create personal geodatabase* e *Create feature dataset*. Na primeira deve-se indicar a localização e o nome outorgado à base de dados geográfica (BDG) e, na segunda, para além do nome, designa-se o sistema de referência espacial, sendo o PT-TM06/ETRS89. Assim, o *layer* de fotos de visitantes, originalmente no sistema de referência WGS84, foi convertido (projectado) para o sistema de referência de coordenadas cartesianas PT-TM06/ETRS89.

Por fim, o referido *layer* foi importado para a BDG como uma *feature class*<sup>48</sup>. A partir deste ponto todos os dados são armazenados na BDG (dados\_lx.gdb) (Figura 4.13).

---

<sup>47</sup> Uma *feature dataset* é um conjunto de *feature classes* relacionadas entre si, que possuem o mesmo sistema de coordenadas.

<sup>48</sup> Uma *feature class* é um conjunto homogéneo de entidades comuns, cada uma com a mesma representação espacial - mesmo tipo de geometria - e um conjunto comum de atributos.



**Figura 4.13: Dados para o estudo organizados na base de dados espaciais.**

## **5. Análise da distribuição espacial dos turistas em Lisboa**

### **5.1. Pré-processamento dos dados espaciais**

Os dados de base utilizados neste estudo foram: i) as *geotagged* fotos dos visitantes, representadas espacialmente num *layer* de pontos; e ii) a divisão administrativa das freguesias de Lisboa, procedente da CAOP 2015. Da CAOP utilizou-se a divisão administrativa das freguesias mas como os limites das freguesias ribeirinhas se estendem até ao meio do rio, tomou-se como referência o limite administrativo da cidade da CAOP 2012 para delimitar as áreas destas freguesias pela linha de costa. A partir desta última operação, foi gerado um novo *layer* com a área limítrofe da cidade.

Com o intuito de simplificar o processamento computacional requerido pelos métodos de análise espacial e facilitar a leitura dos resultados, as *geotagged* fotos foram agregadas em unidades de área. Para a etapa de exploração dos dados, como se pretendia ter uma ideia geral da distribuição espacial dos turistas, os dados foram agregados ao nível das freguesias. Para a etapa de análise espacial, considerando a variação do processo espacial ao longo do território, as fotos foram agregadas em unidades de área hexagonais, de menor dimensão que as das freguesias. Os procedimentos aplicados, em ambiente SIG, para a preparação dos dados são apresentados nas seguintes secções.

#### **5.1.1. Delimitação da área de estudo**

Para a delimitação da área de estudo, executou-se uma operação de selecção por atributos (*Select by attributes*), identificando-se exclusivamente as 24 freguesias de Lisboa, e criando-se um novo tema, apenas com essa informação. Logo de seguida, através da ferramenta *dissolve*, agregaram-se todas as áreas adjacentes (das freguesias) num só polígono que replica a extensão do município.

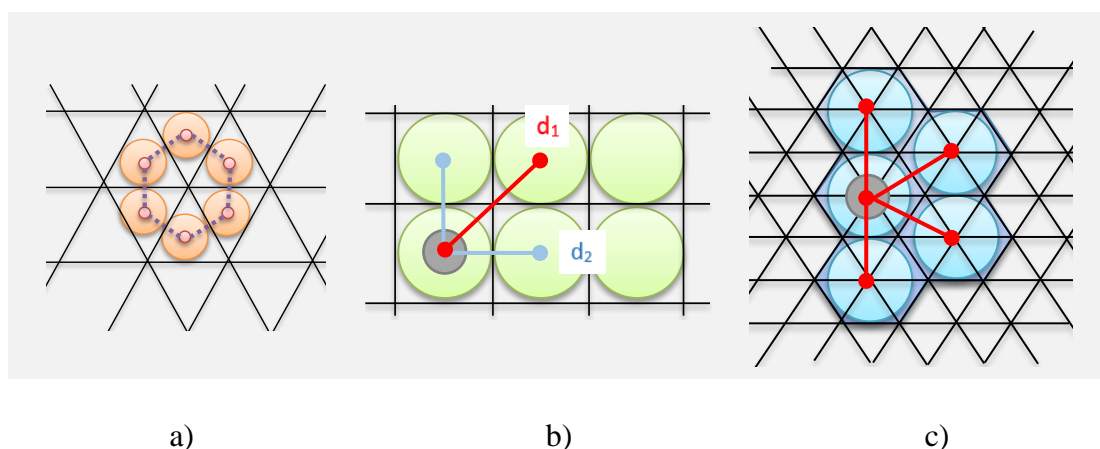
#### **5.1.2. Agregação das fotos por freguesias**

A agregação por freguesias tem por base funções de sobreposição e inquirição espacial entre os temas das fotos e da divisão administrativa das freguesias. Para cruzar a informação de ambos os *layers* e contar as fotografias localizadas dentro de cada área (das freguesias), aplicou-se uma operação de união espacial (*Spatial Join*), de tipo um-

para-um, designando os polígonos das freguesias como *target areas*. Os mapas resultantes são apresentados na secção 5.2.4.

### 5.1.3. Agregação das fotos por unidades hexagonais

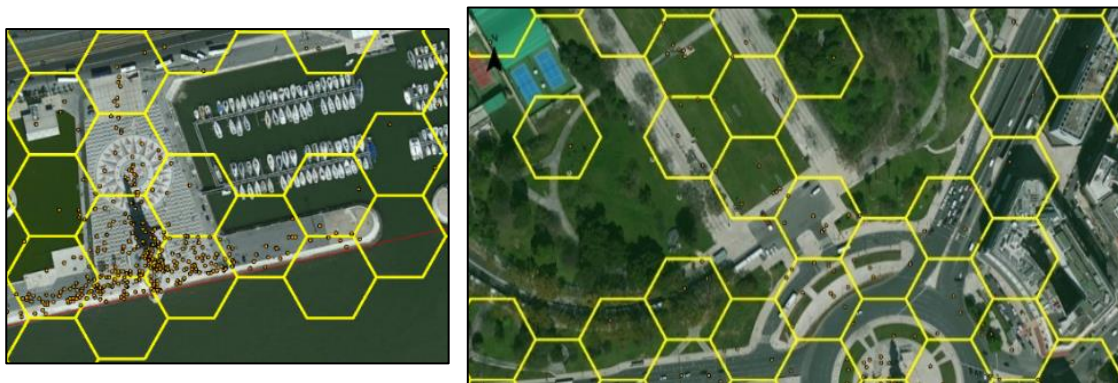
No que respeita à geometria das células da malha regular a utilizar no trabalho, existem três polígonos regulares capazes de preencher o território com unidades de igual área: triângulos, quadrados e hexágonos (Figura 5.1). Das três possibilidades levantadas, a forma mais eficiente é através do hexágono. Quando se analisa a proximidade entre uma célula e as suas vizinhas, no caso de uma malha regular com células quadradas ou triangulares, nem todas as células de uma vizinhança apresentam a mesma distância à célula central (na Figura 5.1 b, a distância entre os centroides das células situadas na horizontal e na vertical da célula *target* é menor do que a distância das células situadas na diagonal,  $d_1 > d_2$ ). No caso da malha regular com células hexagonais, a distância entre os centroides das células numa vizinhança e a célula central é igual em todos os casos (Figura 5.1 c).



**Figura 5.1: Polígonos utilizadas para a criação de malhas regulares.**

Relativamente à área da célula, procurou-se uma dimensão que conseguisse discernir a irregularidade da distribuição espacial das fotos, mesmo nos núcleos onde a frequência da amostra era elevada. Logo, depois de algumas tentativas, determinou-se que a dimensão da célula a utilizar seria de 2.346m<sup>2</sup> metros, pois ajustava-se com mais acerto à disposição das fotos (Figura 5.2). Recorde-se que devido à UAM, qualquer alteração nas características da célula pode produzir resultados diferentes (Rodrigues e Tenedório, 2016; Openshaw, 1983).



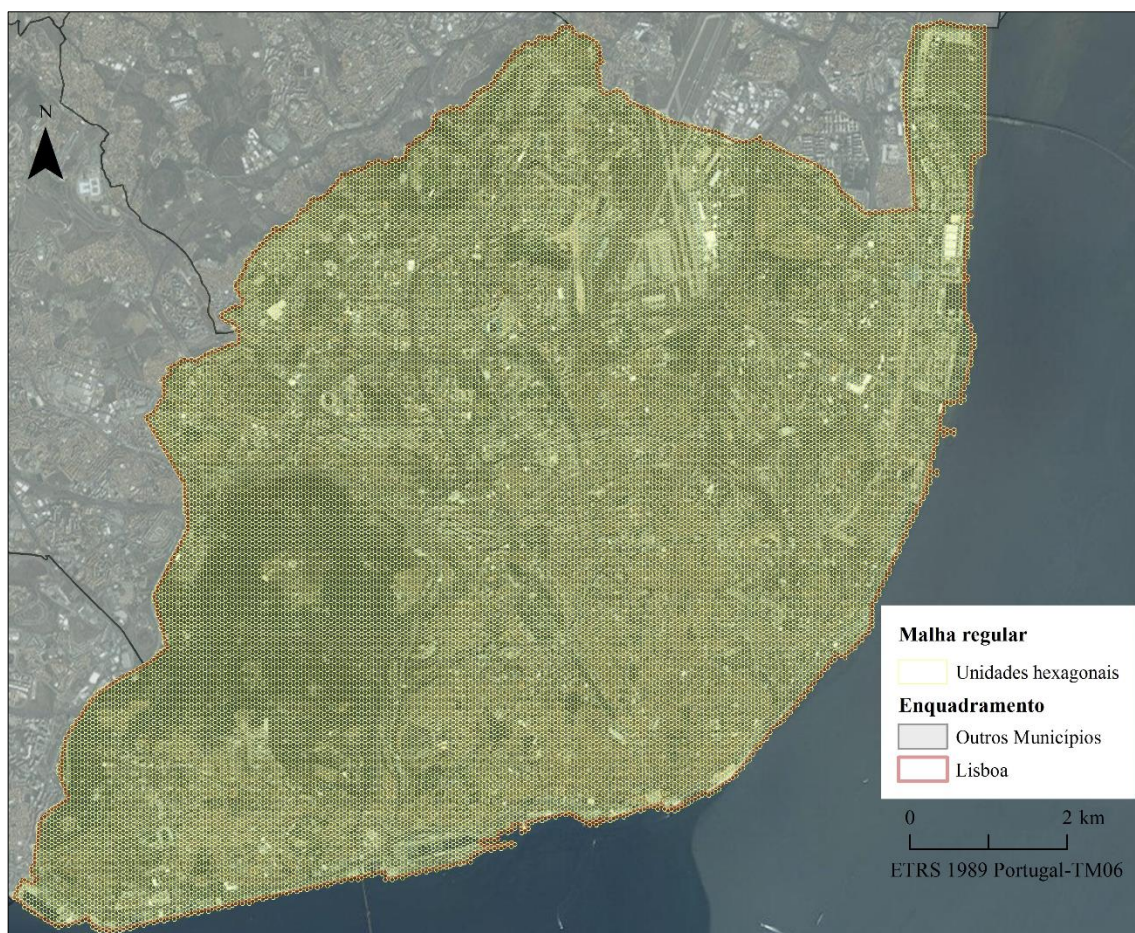


**Figura 5.2: Dimensão da célula da malha regular, agregação por hexágonos.**

Para efectuar a agregação dos dados, foi necessário, em primeiro lugar, criar uma malha regular com células hexagonais. Não havendo uma ferramenta disponível na *toolbox* do ArcMap para executar esta operação, descarregou-se do repositório do ArcGIS (ver Tabela 4.4) a ferramenta *Create Hexagon Tessellation*. Após a sua instalação e importação para a *toolbox* do ArcMap, a ferramenta fica disponível no catálogo de ferramentas.

Para produzir a malha, deve-se indicar a área de referência onde o novo *layer* vai ser criado e a dimensão da célula. Para que a malha se alinhe com a localização da área trabalho tomou-se como referência o *layer* que continha a delimitação da cidade de Lisboa (secção 5.1.1.). Relativamente à dimensão da célula, neste trabalho, optou-se por uma área  $\approx 2.346\text{m}^2$  que corresponde à área média de uma subsecção.

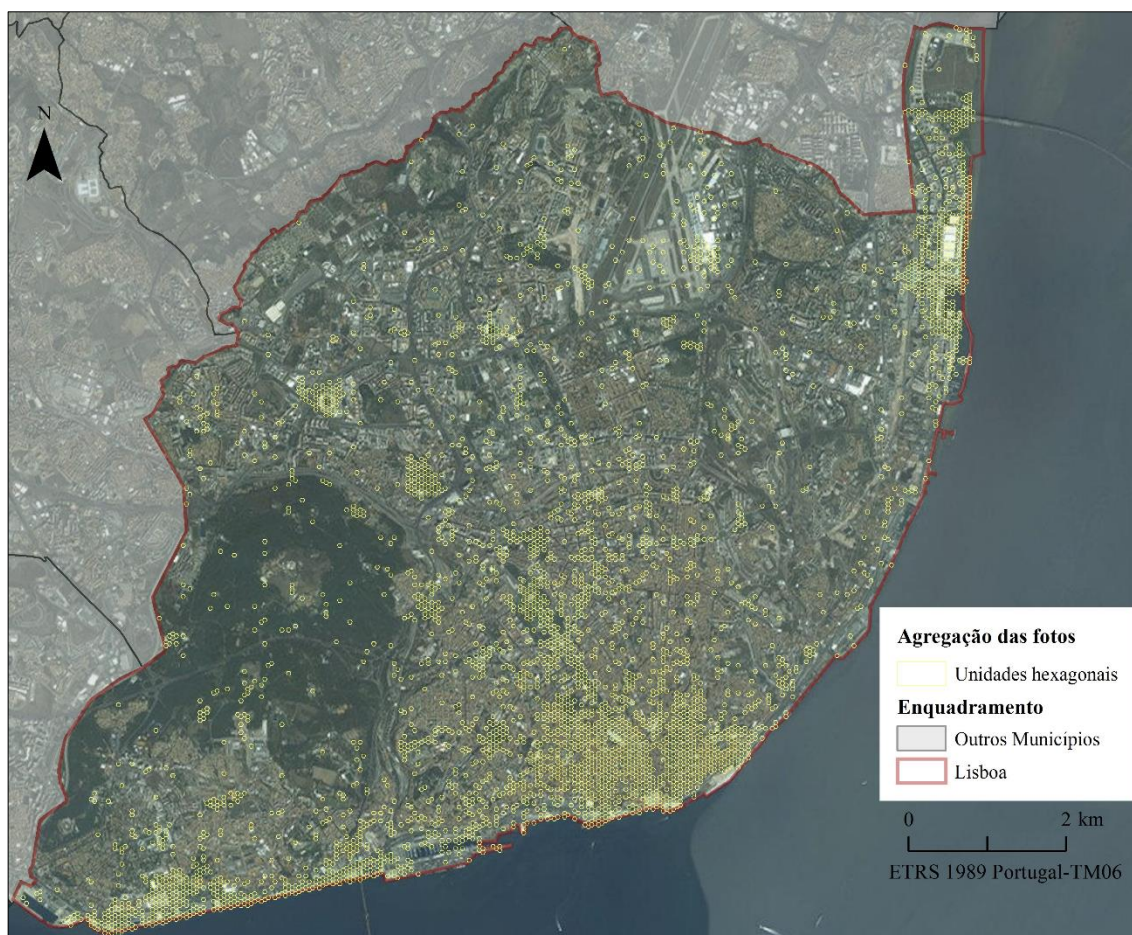
Após a execução da ferramenta, cria-se uma nova camada de informação com a malha regular de hexágonos. Como se pode observar no Anexo 3, a ferramenta criou também unidades (células) fora da área que corresponde à cidade de Lisboa. Para ajustar a extensão da malha à área de estudo, seleccionaram-se unicamente as células situadas sobre a área (Figura 5.3). Assim sendo, ambos os *layers* (a malha de hexágonos e a área de estudo) foram sobrepostos e aplicou-se uma operação de selecção por localização (*Select by location*).



**Figura 5.3: Malha de hexágonos sobre a área de estudo.**

À semelhança do apresentado na secção 5.1.2., para a agregação das fotos por unidades hexagonais, efectuou-se uma contagem das fotos que coincidiam com as células da malha. No novo *layer*, só permanecem as unidades com pelo menos uma foto (Figura 5.4). O total de fotos em cada célula pode ser consultado no campo “Join count” (que é criado automaticamente) da tabela de atributos associada ao *layer*.





**Figura 5.4: Unidades hexagonais resultantes da agregação das fotos.**

## 5.2. Exploração dos dados

Para a representação gráfica dos dados referentes ao número de fotos por freguesias, estes foram classificados de forma a reflectir a sua variação. Na determinação das classes, utilizaram-se os métodos de classificação disponíveis no *ArcGIS*, classes naturais (diagrama de dispersão) e desvio padrão. Após a classificação dos dados, geraram-se dois mapas temáticos apresentados na secção 5.2.4.

Nesta fase, também é apresentado o mapa que mostra a localização geográfica das fotos dos turistas. Em complemento ao mapa, descreve-se a distribuição espacial da amostra, identificando quais os locais onde a concentração de fotografias é maior (secção 5.2.1). Com o intuito de perceber o grau de concentração (em termos absolutos), computou-se o número de fotos associadas a cada um desses locais. Deve-se notar que os valores apurados são apenas aproximações, uma vez que, na realidade, a área de

influência do elemento de referência pode ser diferente da delimitada segundo a metodologia proposta.

Assim, para determinar a quantidade de fotos associadas aos pontos de interesse, efectuou-se uma contagem das fotos tiradas dentro de um raio de vizinhança de até 20m em torno do local (cf. Tabela 5.1e Tabela 5.2). As áreas de influência dos locais foram elaboradas tendo como base a informação vectorial referente à rede viária disponível no *Open Street Map*<sup>49</sup> (as áreas são ilustradas da Figura 5.5 à Figura 5.12). Estas áreas de influência foram subdivididas em duas zonas de proximidade de 10m e 20m, utilizando a ferramenta *Buffer*. Para a contagem das fotos, aplicou-se uma operação de união espacial, como explicado no ponto referente à agregação dos dados (secções 5.1.2 e 5.1.3).

Ainda relativamente à amostra das fotografias, na secção 5.2.3., são apresentados algumas características temporais dos dados. Para traçar a evolução do número de fotografias e de visitantes, segmentaram-se os dados por meses e por anos, segundo a data em que foram carregadas.

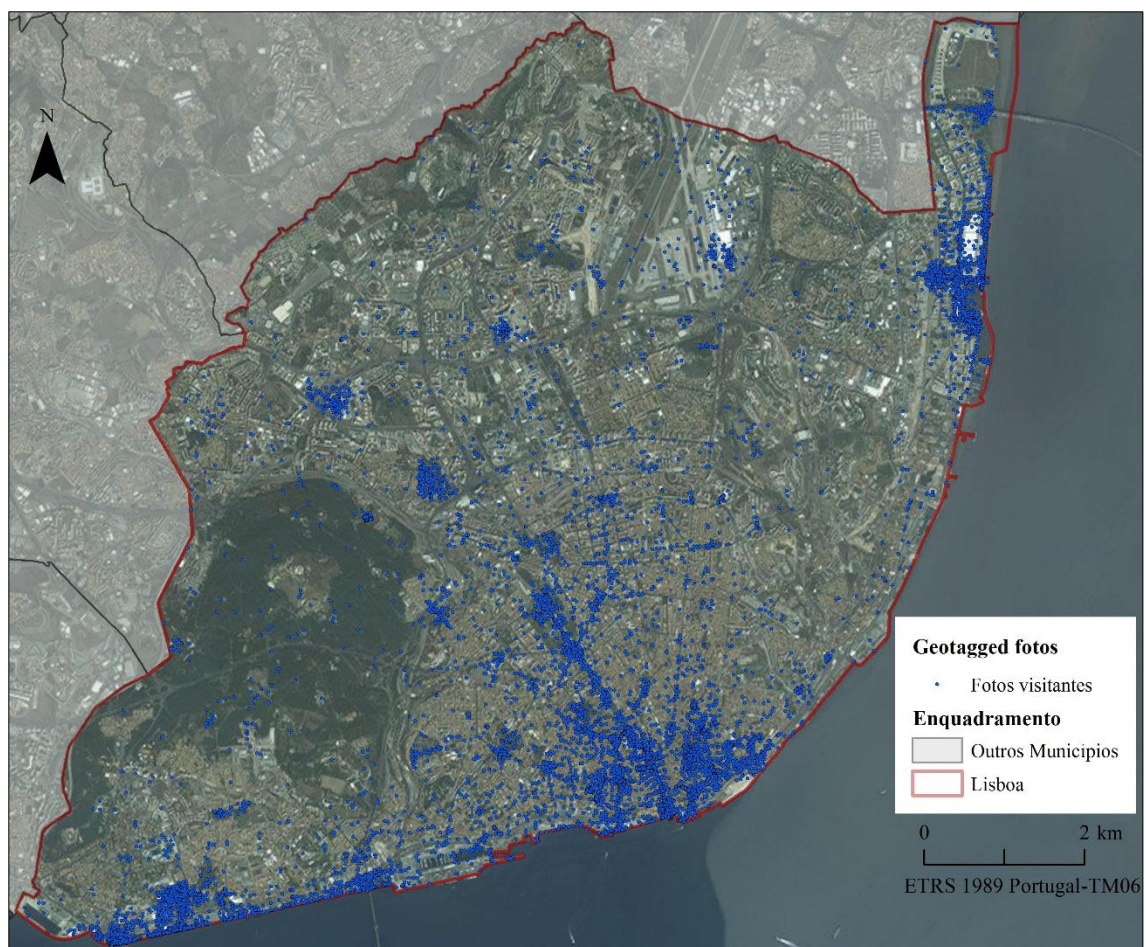
A análise dos mapas enquadrada nesta secção deve ser entendida como um exercício preliminar à etapa de análise estatística que procura, meramente, abrir a discussão sobre a distribuição espacial dos turistas. A análise de padrões é tratada na secção 5.3.

#### **5.2.1. A localização das fotografias dos turistas**

A localização das fotografias tiradas pelos turistas é observável na Figura 5.5. Da leitura ao mapa denota-se, à primeira vista, que existem áreas com maior densidade de fotos claramente distinguíveis das restantes.

---

<sup>49</sup> O *Open Street Map* é um mapa de licença aberta, desenvolvido por uma comunidade voluntária de mapeadores que contribuem e mantêm atualizados os dados sobre estradas, trilhos, restaurantes, estações de metro e autocarros, e outros objectos interesse. Em *Openstreetmap*. url: <https://www.openstreetmap.org/about>



**Figura 5.5: Fotos dos turistas de Lisboa.**

No nordeste da cidade, a zona do Parque das Nações (Figura 5.6) aparece como um local isolado de elevada concentração, com poucas fotos repartidas pela sua vizinhança. Dentro do parque, destacam-se o Oceanário, os exteriores do Centro Comercial Vasco da Gama e da Estação Oriente, o passeio das Tágides e, no ponto mais para o norte, a Torre Vasco da Gama. Para o interior da cidade, surge um outro foco de fotografias, em torno da terminal 1 do Aeroporto de Lisboa, com as fotos na sua maioria concentradas nos espaços próximos à fachada do edifício e nas pistas de aterragem.





**Figura 5.6: Locais de interesse, nordeste da cidade (Parque das Nações).**

O centro da cidade mostra-se como uma área muito fotografada, com alguns núcleos de maior relevância (Figura 5.7). A disposição das fotos formam um eixo central perpendicular à margem do rio, que vai da Praça de Comercio ao interior, passando pela rua Augusta, Praça dos Restauradores, Av. da Liberdade, e Marquês de Pombal, até o Parque Eduardo VII. No início do eixo, sobressai a Praça de Comércio. A maioria das fotos convergem à volta da estátua de D. José I e no Cais das Colunas. Na rua Augusta, a concentração de fotos é superior no traçado desde o Arco até a rua São Nicolau, diminuindo conforme se avança para o extremo norte. Nas ruas paralelas, apesar de serem passeios pedonais (como a rua Augusta), a presença de imagens é quase nula. Ao contrário, na rua transversal Santa Justa, desde a rua Augusta até à rua do Carmo (sentido Este–Oeste), o número de imagens é semelhante ao da rua Augusta, sendo mais intenso em redor do Elevador de Santa Justa. Na mesma direcção, nota-se uma acumulação de imagens no Largo do Carmo.





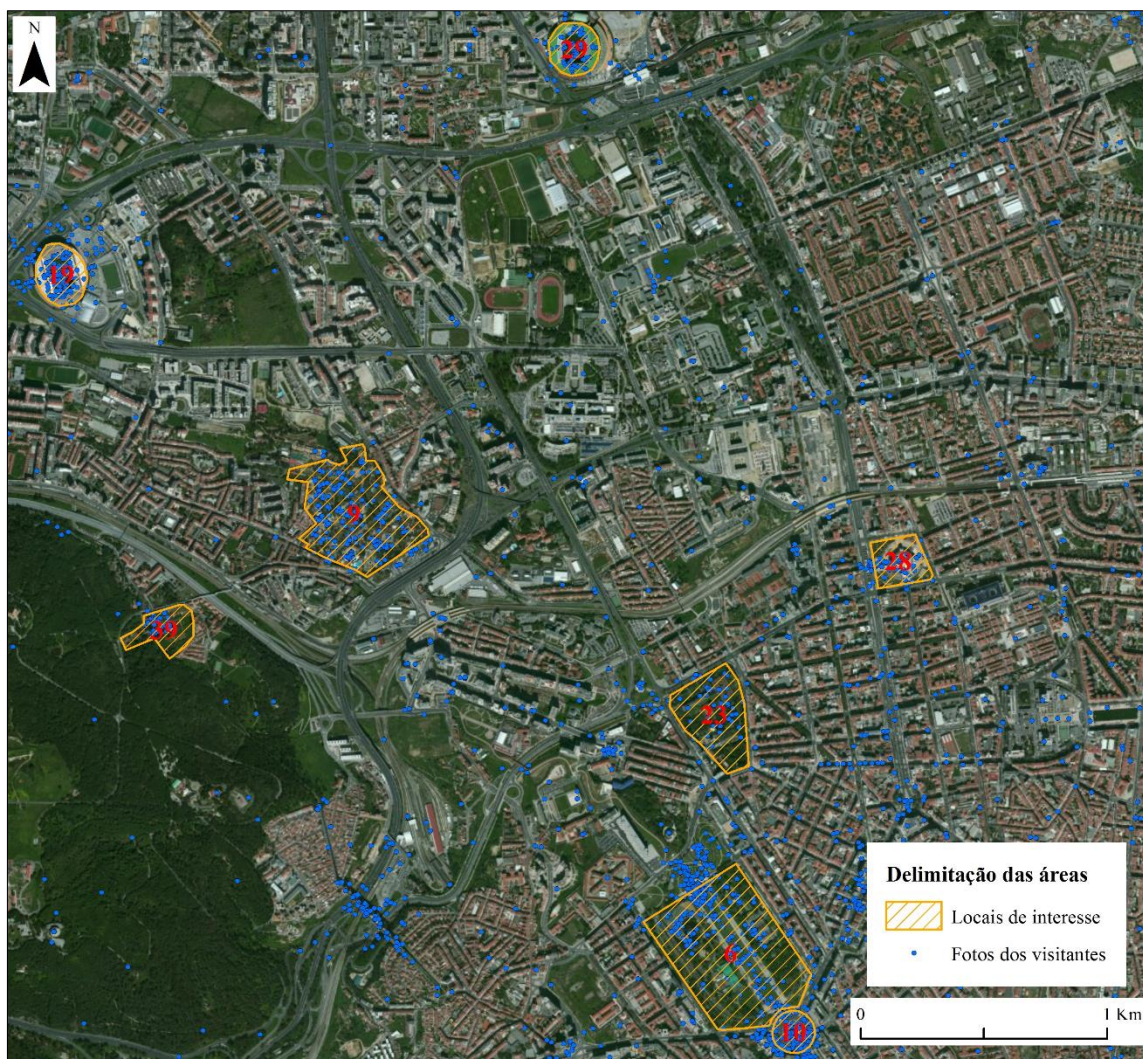
**Figura 5.7: Locais de interesse, eixo central (Praça de Comercio – Praça dos Restauradores).**

A seguir, as praças Dom Pedro IV e dos Restauradores aparecem como dois focos activos de fotografias, a última, com mais do dobro de fotos em comparação com a primeira. Ao longo do trajecto que conecta estas duas praças, observam-se algumas ocorrências sobretudo instaladas em frente da estação do Rossio. Nas praças da Figueira e Martim Moniz, a quantidade de fotos é menor do que nas praças antes referidas. As fotos ao longo da Av. Liberdade, entre a Praça dos Restauradores e o Marquês de Pombal, concentram-se estritamente no próprio eixo. Com excepção das fotos associadas ao percurso do Elevador da Glória (com maior intensidade nas paragens, em ambos os



extremos, do que no trajecto em si), ao miradouro de São Pedro de Alcântara e ao jardim Botânico de Lisboa, nas faixas paralelas à Av. Liberdade, a presença de fotos é mínima, diminuindo à medida que nos afastamos do eixo.

Olhando com mais pormenor, as fotos posicionadas na parte final do eixo (no parque Eduardo VII) agrupam-se, particularmente, em dois locais: no sul junto ao monumento do Marquês de Pombal, e no lado oposto, junto ao jardim Amália Rodrigues (Figura 5.8).



**Figura 5.8: Locais de interesse, eixo central (Marquês de Pombal), e Interior.**

Mais para o interior, encontram-se vários locais pontuais de menor relevância. Estes são, sem nenhuma ordenação específica: o jardim da Fundação Calouste Gulbenkian (incluindo o Museu da Fundação e o Museu de Arte Moderna), a Praça de Touros do Campo Pequeno e o Palácio dos Marquesses de Fronteira. Continuando neste sentido, descobrem-se ainda três locais, com mais fotos do que o grupo anterior, são eles



o jardim Zoológico de Lisboa e os estádios da Luz e José Alvalade (destas duas infra-estruturas desportivas, a primeira apresenta um maior número de fotos do que a segunda).

Continuando a acompanhar o eixo central (Figura 5.9), mais próximo da margem do rio, distinguem-se dois núcleos a Nascente e a Poente, nos bairros de Castelo e Alfama; e da Baixa-Chiado e Bairro Alto, respectivamente.



**Figura 5.9: Locais de interesse, núcleo centro-este.**

De salientar, no lado Este, o Castelo de São Jorge onde se verifica uma quantidade importante de imagens, em especial, na entrada da fortaleza (na parte da planta rectangular) e no miradouro do castelo. Em Alfama, são os monumentos religiosos e os miradouros, os que atraem mais fotografias. O acumulado de fotos cobre os miradouros das Portas de Sol e de Santa Luzia, e os espaços imediatos às Igrejas Sé de Lisboa, São



Vicente de Fora, e ao Panteão Nacional (Igreja de Santa Engrácia). Na Graça, constata-se apenas um foco de imagens agregadas junto à Igreja da Graça e ao jardim da Cerca da Graça.

No lado Oeste do eixo, o conglomerado de fotos revela o valor turístico de dois locais, a Praça Luís de Camões e o percurso do Elevador da Bica (Figura 5.10). O número de fotos sobre a praça é complementado pelas fotos ancoradas ao Largo do Chiado. Em sentido Este–Oeste, seguindo a rua de Loreto, emerge um outro foco centrado na estação norte do Elevador. O raio de influência deste foco estende-se, com menor intensidade, para o Sul (sobre o percurso do Elevador). Na freguesia de Santa Catarina, à semelhança do que acontece na Graça, constata-se um único ponto de interesse, o miradouro de Santa Catarina. Para nascente, tendo como referencia a rua Calçada da Estrela, identificam-se dois grupos de imagens, situados na Assembleia da República e no convento do Santíssimo Coração de Jesus. Em menor proporção, identificam-se algumas fotografias no jardim da Estrela e no Cemitério dos Prazeres.



**Figura 5.10: Locais de interesse, núcleo centro – oeste.**

No Sul da cidade (Figura 5.11), a concentração de fotos estende-se, principalmente, pela margem ribeirinha, desde o Cais de Sodré, passando por Alcântara até Belém. Existe uma diferença na quantidade de fotos enquadradas na primeira e segunda parte deste eixo, a favor do troço Sul (de Alcântara à Belém). No troço que vai do Cais de Sodré a Alcântara (Doca de Alcântara), as fotos localizam-se nas proximidades

da zona de bares e discotecas (em Santos) e na doca do Espanhol. No segundo troço, o arranjo das fotografias seguem a forma do passeio pedonal entre a doca de Alcântara e a doca de Belém (a fracção de fotos intensifica-se com o aumento da proximidade a Belém). Em ambos os troços é notório que as imagens tendem a permanecer perto da margem.



**Figura 5.11: Locais de interesse, eixo ribeirinho (Tramo Alcântara - Belém).**

Em Belém (Figura 5.12), os locais com maior número de fotografias estão relacionados com os monumentos: Padrão dos Descobrimentos, Torre de Belém, e Mosteiro dos Jerónimos. O volume de fotos no Padrão dos Descobrimentos é menor do que na Torre de Belém, e a mancha de pontos que envolve o primeiro é também menos extensa do que a segunda. As imagens associadas ao Mosteiro assentam na sua parte externa (imediate à fachada) e no jardim interior. Um pouco mais dispersas, dentro do “complexo turístico”, detectam-se unidades posicionadas, sobretudo, nas praças, esplanadas e nos espaços verdes conexos às atracções. Identificam-se outros focos de menor relevância (quando comparado com os monumentos), nomeadamente o Centro Cultural de Belém, os museus de Arte Popular, dos Coches, da Marinha, da Colecção Berardo, o museu e o monumento aos Combatentes, e o Palácio de Belém. Nesta mesma parte da cidade, mas mais para o interior, surge uma pequena concentração de imagens no jardim Botânico da Ajuda.





**Figura 5.12: Locais de interesse, núcleo sul (Belém).**

A Torre de Belém e a Praça de Comércio são, de longe, as atracções que mais se destacam (Tabela 5.1). A quantidade de fotos associadas a estes locais é muito elevada, mesmo quando comparada com outros locais turísticos da cidade. Todavia, se agruparmos ambos os valores resultantes da contagem, os dois locais, em conjunto, representam pouco mais do 13% do total de fotografias (dos turistas).

Esta situação não surpreende, pois é evidente a presença de fotos segregadas fora das zonas turísticas da cidade (no exterior do contorno das microcentralidades do turismo em Lisboa). Se olharmos para a distribuição espacial destas fotos, esta tende a ser mais dispersa do que das fotos localizadas dentro das zonas turísticas e, como tal, não se registam focos de aglomeração dignos de menção. Naturalmente, sendo esta uma amostra de fotos geradas por turistas, nos sectores residenciais, principalmente no lado Nordeste da cidade, existem locais com total ausência de registos.

**Tabela 5.1: Número de fotos registadas sobre os locais de interesse - atracções turísticas (por questões de funcionalidades apenas estão representados os locais com mais de 45 fotos).**

Nº	Local de interesse	Fotos dos turistas			
		Dentro da área	Área de influência 10m	Área de influência 20m	Total
1	Torre de Belém	1.056	90	137	1.283
2	Praça do Comércio	960	56	108	1.124
3	Castelo de São Jorge	677	51	79	807
4	Mosteiro dos Jerónimos	562	93	146	801
5	Centro Comercial Vasco da Gama	357	36	86	479
6	Parque Eduardo VII	336	43	67	446
7	Padrão dos Descobrimentos	252	59	84	395
8	Oceanário de Lisboa	265	8	15	288
9	Jardim Zoológico de Lisboa	228	5	11	244
10	Marquês de Pombal	147	20	40	207
11	Miradouro de São Pedro de Alcântara	116	20	65	201
12	Praça Dom Pedro IV (Rossio)	146	17	36	199
13	Praça Luís de Camões	66	42	80	188
14	Praça dos Restauradores	70	36	74	180
15	Percurso do Elevador da Bica	-	179*	-	179
16	Rua de Santa Justa – Elevador de Santa Justa	-	168*	-	168
17	Igreja da Graça	75	33	59	167
18	Rua Augusta	-	160*	-	160
19	Estádio da Luz	138	6	13	157
20	Praça do Império	101	16	28	145
21	Percurso Elevador da Glória	-	143*	-	143
22	Igreja da Sé - Largo da Sé	67	32	42	141
23	Fundação Calouste Gulbenkian	46	3	88	137
24	Centro Cultural de Belém	84	12	23	119
25	Passeio das Tágides - Parque das Nações	-	114*	-	114
26	Eixo Ribeirinho (Alcântara - Belém)	-	92*	-	92
27	Miradouro das Portas do Sol	34	17	40	91
28	Praça de Touros do Campo Pequeno	46	15	27	88
29	Estádio José de Alvalade	76	3	6	85
30	Panteão Nacional	66	2	9	77
31	Museu do Combatente	66	3	7	76
32	Miradouro de Santa Catarina	40	12	20	72

\* Estes elementos correspondem a passeios ou percursos. Considerando a configuração geométrica das áreas destes locais, de tipo linear, para o cálculo do número de fotografias, abarcou-se apenas uma área de influência de 10m para ambos os lados do traçado original.

Nº	Local de interesse	Fotos dos turistas			
		Dentro da área	Área de influência 10m	Área de influência 20m	Total
33	Torre Vasco da Gama	29	11	20	60
34	Miradouro de Santa Luzia	6	19	27	52
35	Igreja de São Vicente de Fora	35	6	11	52
36	Assembleia da República	29	5	17	51
37	Jardim Botânico de Lisboa	46	1	3	50
38	Convento Santíssimo Coração de Jesus	30	6	11	47
39	Palácio dos Marquês de Fronteira	45	0	1	46

Verifica-se ainda que, com excepção dos dois estádios de futebol (Luz e José Alvalade) e do palácio Marquês de Fronteira, a maioria dos sítios são abrangidos pelas microcentralidades de turismo, e que os focos de maior aglomeração de fotos convergem para dentro das chamadas zonas turísticas da cidade.

Por outro lado, as diferenças de grau de dispersão entre as fotos dos turistas e locais pode ser percebida através da comparação dos valores da distância padrão (*standard distance*<sup>50</sup>) para cada conjunto de dados. Como se denota nos resultados (Tabela 5.2, Anexo 4 e Anexo 5), as fotografias dos turistas possuem uma distância padrão inferior às dos locais. Ambos os centros médios diferem na sua localização, estando o dos locais posicionado ligeiramente mais para o interior, a uma distância de 113,98m face ao centro médio dos turistas.

**Tabela 5.2: Valores da distância padrão das fotografias dos turistas e locais.**

Conjunto de dados	Distância padrão	Coordenadas centros médios
Fotos de turistas	3.837,1	X: -88739,37; Y: -104643,05
Fotos de locais	4.254,5m	X: -88694,51; Y: -104538,26

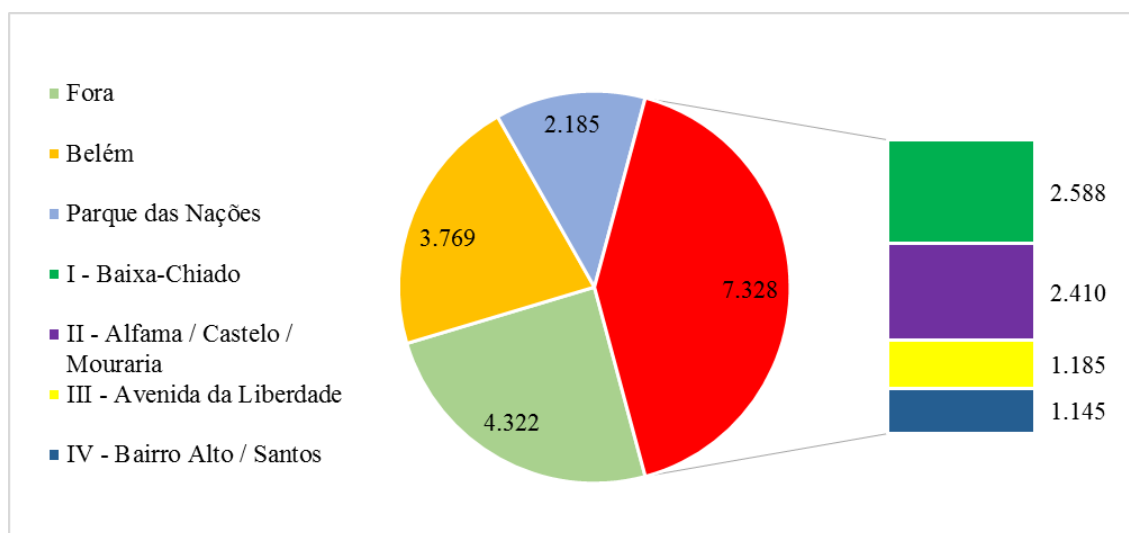
### 5.2.2. Distribuição das fotos dos turistas por microcentralidade

Tendo em atenção as microcentralidades definidas segundo o zonamento proposto no PETRL 2015 – 2019 (Figura 4.2) pode-se ter uma ideia das preferências dos turistas relativamente às zonas da cidade, consideradas como áreas de aposta prioritária (Figura

<sup>50</sup> *Standard distance* ou distância padrão, mede quanto variam as distâncias entre o centro médio e as entidades, da distância média observada. Quanto maior for o resultado maior será a dispersão das *features* (Mitchell, 2005).

5.13). Efectivamente, 75% das fotos são tiradas dentro das áreas definidas como microcentralidades. Este valor divide-se entre o centro histórico, principal foco de atracção com mais de 40% do número total de fotos, Belém (cerca de 20%) e Parque das Nações (com 12%).

Apesar da predominância do centro histórico, a distribuição das fotografias dentro dos seus limites é geograficamente assimétrica, podendo claramente observar-se a predominância de duas subáreas sobre as demais. Neste contexto, as subáreas I (Baixa-Chiado) e II (Alfama / Castelo / Mouraria) com, respectivamente, 35% e 32% do total de fotos tiradas no centro histórico, demarcam-se das restantes, a subárea III (Avenida da Liberdade) com 16% e a IV (Bairro Alto / Santos) com 15%.

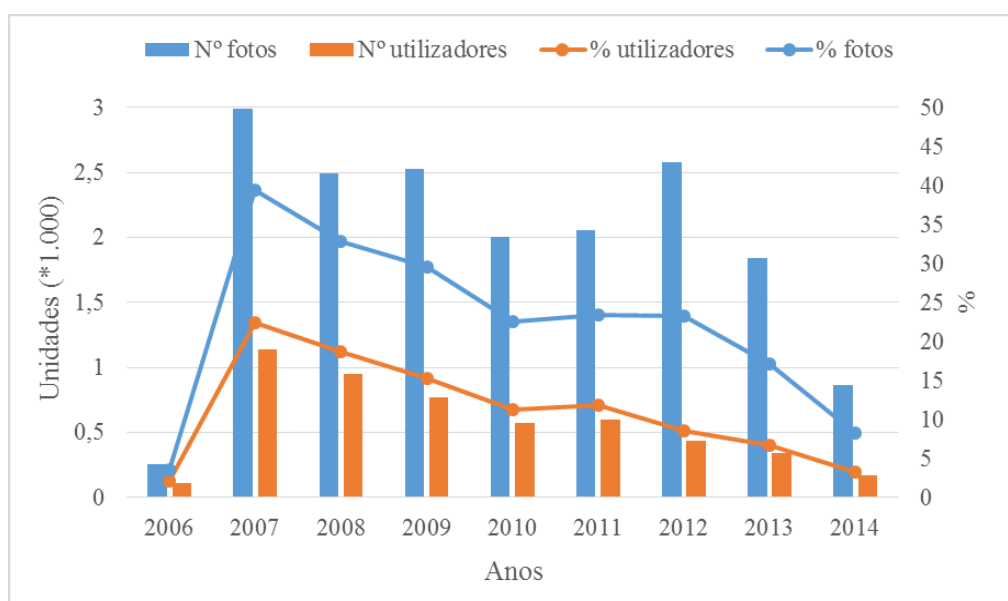


**Figura 5.13: Distribuição das fotografias pelas microcentralidades.**

### **5.2.3. Classificação de fotos segundo a data em que foram carregadas**

Quando se passa à análise temporal dos dados, sobressai a tendência decrescente do volume de fotografias (de 2.988 no ano 2007 a 862 no ano 2014), ainda mais, quando no último ano a queda é de aproximadamente 1.000 unidades. O decréscimo no número de fotografias entre os dois últimos períodos pode estar relacionada com o facto de não se ter conseguido extrair fotos em todos os meses do ano 2014. Apesar das várias tentativas de extracção, executadas no dia da descarga de informação (15 de Dezembro, 2015), os registos mais recentes, nos dados originais (locais e visitantes), eram referentes ao mês de Agosto 2014, sendo a foto mais recente datada do dia 11 desse mês. Não foi possível obter fotografias dos meses seguintes, nem do ano 2015 (Anexo 6).

Assim sendo, muito embora não seja possível confirmar a evolução negativa no último ano, como se pode observar na Figura 5.14 e no Anexo 11, denota-se que a redução do volume da amostra vem desde o ano 2013. Esta mesma tendência é verificada quando observarmos o conjunto original de dados, agregando as fotos dos locais e turistas. De 2006 a 2012 os valores seguem uma tendência positiva, sendo que nos anos de 2013 e de 2014, esta tendência reverte-se, com uma grande diferença de fotos no último ano, de 12.949 para 6.005 fotografias, apesar de influenciada pela falta de dados nos últimos meses.



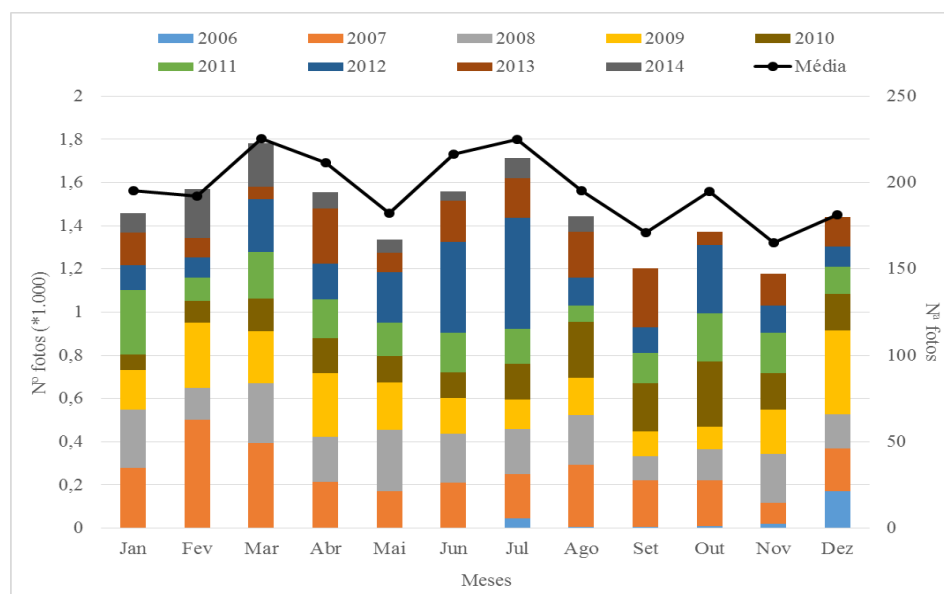
**Figura 5.14: Número de fotos de turistas extraídas do Panoramio, 2006-2014.**

No que respeita aos utilizadores considerados como turistas, com excepção do valor registado no ano 2011, de 2007 a 2013, o número decresce continuamente, sendo a tendência de baixa mais constante do que o verificado nas fotografias. Estes resultados, obviamente, discordam dos “números” da procura turística de Lisboa, e não se devem interpretar como uma redução na quantidade de turistas reais, mas apenas como um decréscimo na proporção de utilizadores que partilharam a sua experiência através desta rede social. Nos dados originais (referentes aos turistas e locais), também se regista uma diminuição constante no número de utilizadores, passando de 1.575 em 2007 para 720 e 355 em 2013 e 2014, respectivamente (Anexo 7).

A repartição das fotografias dos turistas ao longo da série temporal não é regular (Figura 5.15), particularmente nos anos 2007 e 2013, onde os valores se afastam consideravelmente da média anual de 2.355 fotos (a média foi calculada incluindo os valores de 2007 a 2013). As fotografias de 2006 são apenas 258, datadas na sua maioria



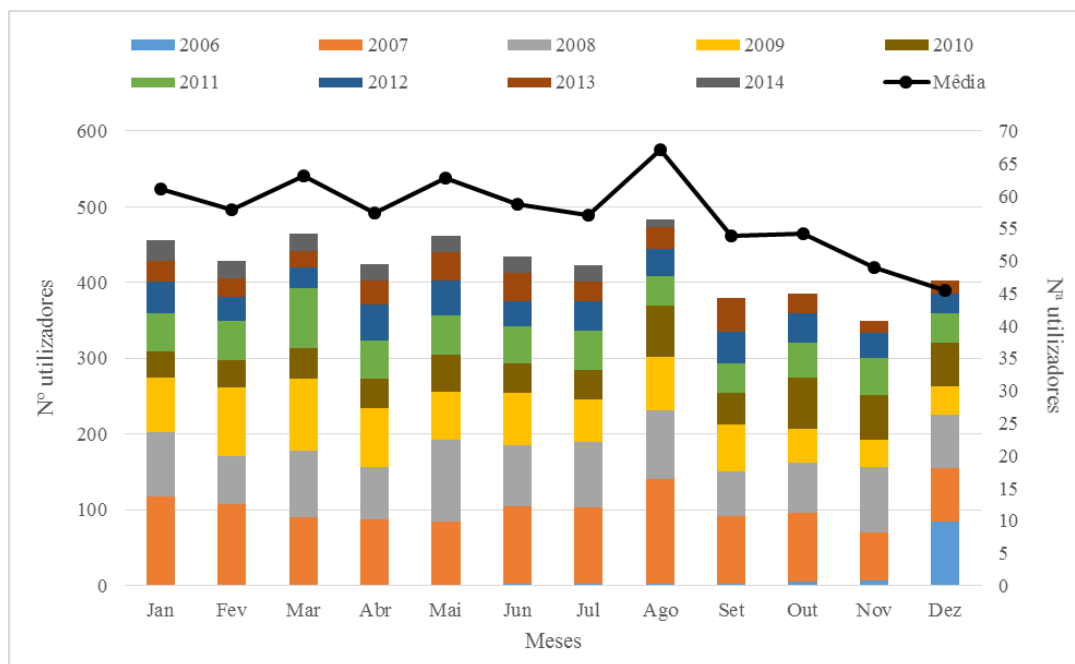
do mês de Dezembro. Este ano só apresenta registros praticamente em 9 meses (em Janeiro, Maio e Junho, o número de visitantes é igual o menor o 2) (Anexo 8). Deste modo, as fotos deste ano não foram consideradas para a análise temporal, pois devido à sua descontinuidade mensal, poderiam enviesar os resultados da análise.



**Figura 5.15: Classificação de fotos dos turistas por meses.**

Em muitos os casos, a evolução mensal do número de fotografias registrada para cada ano é diversa. Os valores para um mesmo mês mudam significativamente de um ano para outro (como são os casos do meses de Fevereiro, Março, Abril, Novembro, Dezembro). O último ano apresenta alguns desníveis quanto à repartição mensal, para além de não existirem registros nos últimos quatro meses, apresenta os valores mais baixos da série de dados nos meses de Abril, Maio e Junho (Anexo 8 e Figura 5.16).

Quanto à evolução mensal do número de utilizadores, esta não é similar à das fotografias. O número de utilizadores mantém-se mais constante, pelo menos nos primeiros sete meses (apresenta valores médios de 57 a 63 utilizadores). Em média, no mês de Agosto, evidencia-se uma subida para 67 utilizadores, para no último mês o valor descer até os 45 (Anexo 9).



**Figura 5.16: Classificação dos utilizadores por meses.**

A tendência registrada no número de turistas que carregam fotografias não difere da sazonalidade turística projectada pelas taxas de ocupação hoteleira. Quando comparados, constata-se subidas nos mesmos meses que nos dados referentes à ocupação. Contudo, olhando para as séries individualmente, verifica-se que o comportamento mensal nos anos de 2012 e 2013 se assemelha mais às estatísticas dos alojamentos, sobretudo nos valores do segundo semestre, do que o verificado nos restantes anos. A correlação entre estas duas variáveis é de  $R^2=0,67$  (tomou-se em consideração os valores médios de cada mês nos dois anos para a taxa de ocupação e o número de utilizadores) o que pode considerar-se satisfatório (Anexo 10).

O rácio de fotos por turista é de 3,8 considerando os valores totais. Quando calculado para cada ano, pode-se verificar que, sensivelmente até 2011, a quantidade de fotos varia de forma semelhante à dos utilizadores, o que não acontece nos anos subsequentes (2012, 2013 e 2014). Nos anos de 2006 a 2011 o rácio fotos/visitante é inferior a 3,8 em contraponto ao verificado nos três últimos anos quando se situa acima dos 5 valores (Anexo 11).

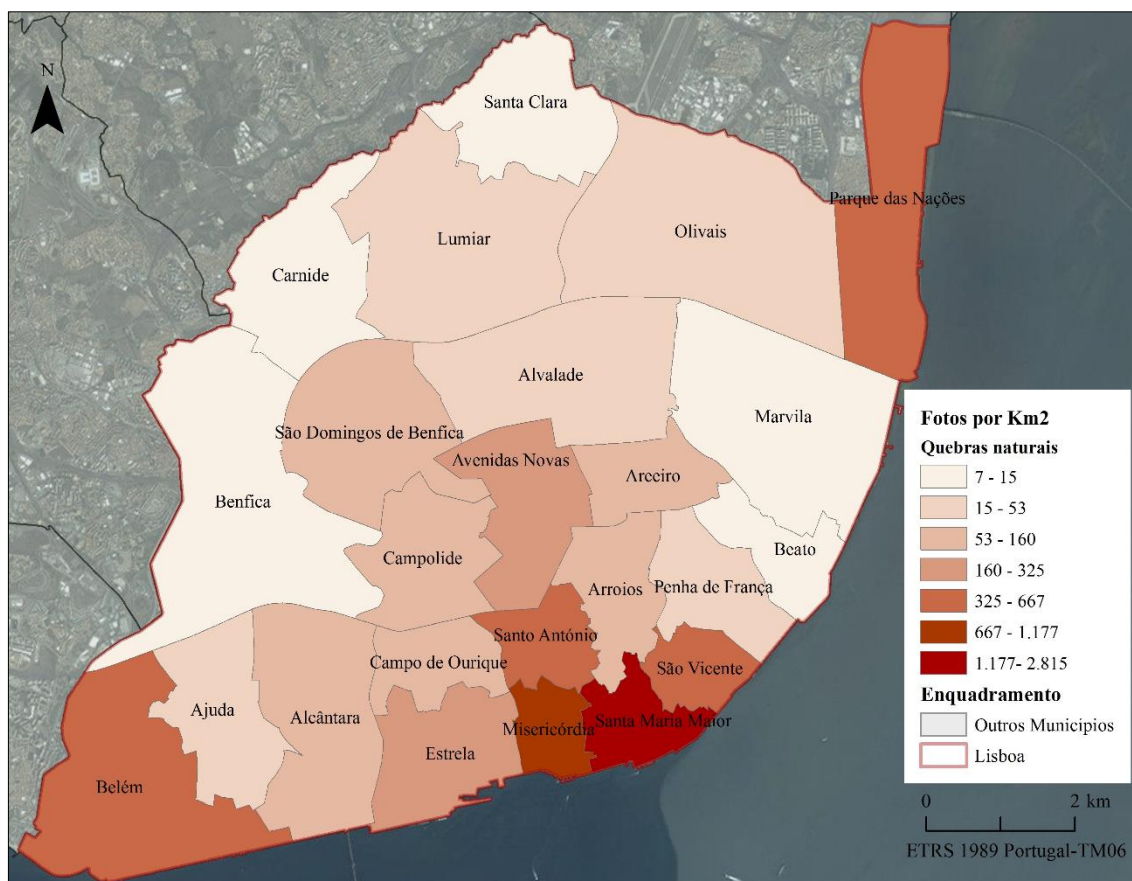
Para encerrar este ponto, importa ressaltar a importância e dimensão da série temporal de amostra, abrangendo oito anos de dados (sem contar com o ano 2006). Contudo deve-se ter em consideração a particularidade do último ano relativamente à ausência de fotografias nos últimos quatro meses.

#### **5.2.4. Distribuição das fotos dos turistas por freguesias**

A amostra de fotografias agregadas por freguesias foi submetida a duas classificações, por intervalos (classes) naturais e por desvio padrão. Antes da classificação, para evitar o enviesamento (*bias*) causado pela disparidade nas dimensões das freguesias, os valores absolutos de número de fotos associados a cada uma foram normalizados pela área (em km<sup>2</sup>) da freguesia em que se inserem.

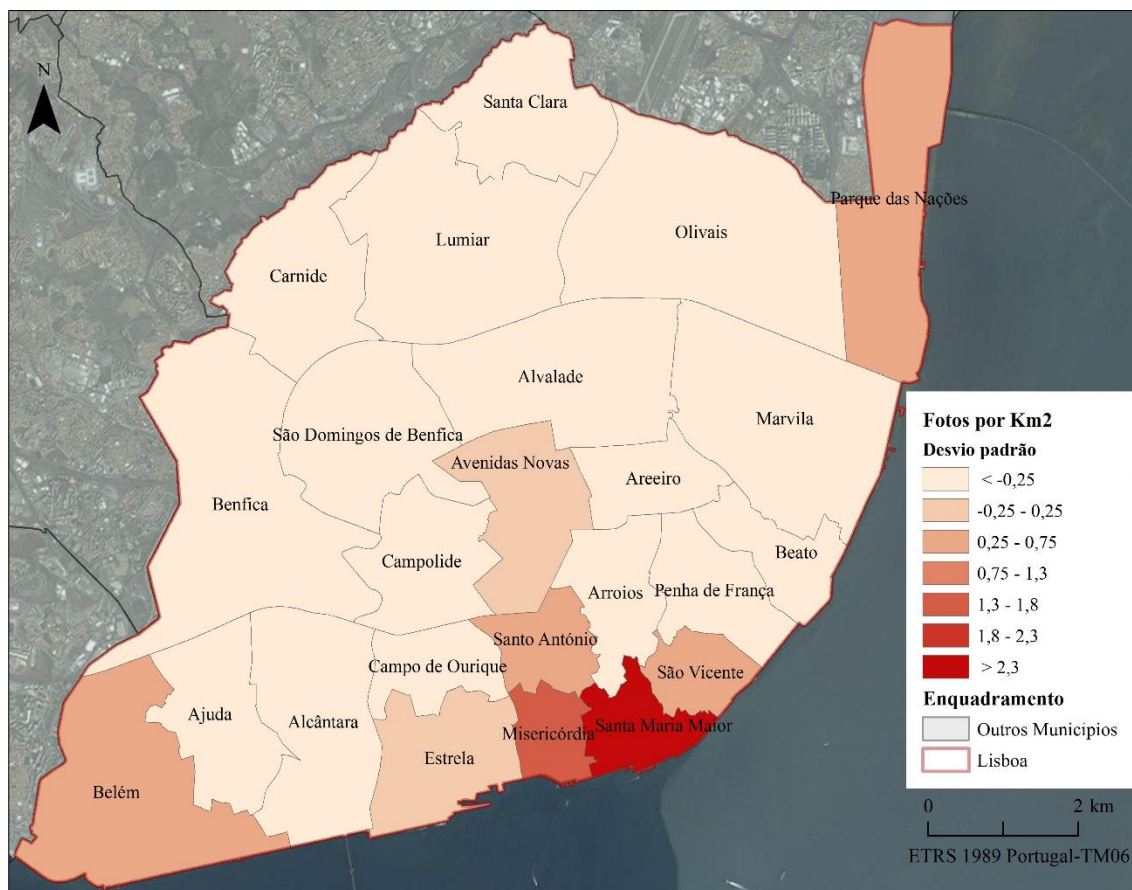
Da leitura cruzada dos mapas da Figura 5.17 e da Figura 5.18, o que ressalva é o progressivo aumento da densidade de fotos, em direcção às freguesias situadas no centro da cidade. Excluindo as freguesias de Marvila e Beato, a tendência positiva é verificada nas restantes freguesias junto ao rio, sobretudo no eixo Sul. Ao contrário, a diminuição na densidade sente-se nas freguesias próximas da fronteira interior do concelho, onde a densidade chega apenas às 29 fotos/km<sup>2</sup>. As freguesias situadas no interior possuem valores de densidades inferiores à média do conjunto de dados, o qual se cifra nas 312,6 fotos/km<sup>2</sup> (ver Anexo 12).

Em concordância com o revisto na secção 5.2.1., o centro da cidade é o sector que apresenta a maior concentração de fotos. De notar que a densidade da freguesia Santa Maria Maior (influenciada pelos locais de interesse, *e.g.* Praça de Comércio e Castelo de São Jorge), se situa muito acima da média, chegando a duplicar os valores da freguesia da Misericórdia. Nas restantes freguesias associadas a este sector, a densidade diminui, passando das 1.177 fotos/km<sup>2</sup> para as 325 fotos/km<sup>2</sup> na freguesia das Avenidas Novas.



**Figura 5.17: Densidade de fotos nas freguesias, classificada pelas quebras naturais.**

No Norte e Sul, emergem dois pólos, que apresentam densidades distintas daquelas registradas nas freguesias vizinhas. No norte, a freguesia do Parque das Nações deve a sua importância às fotos localizadas em torno da infra-estrutura de animação do Parque das Nações (Oceanário, Casino, Torre Vasco da Gama, etc.). No pólo oposto, os ícones turísticos de Belém motivam a proeminência da freguesia com a mesma designação, cuja densidade se encontra 0,75 desvios padrão acima da média, concedendo-lhe ainda mais destaque do que ao Parque das Nações. Em termos absolutos, na freguesia de Belém foram referenciadas mais de 3.800 fotos (Anexo 12).



**Figura 5.18: Densidade de fotos nas freguesias, classificada pelos desvios padrões.**

### 5.3. Análise estatística

Nesta secção são explicitados os métodos estatísticos que suportam a análise de padrões efectuada no estudo. Tendo sido tratada a base conceptual dos métodos na secção 3.3.2, continua-se com a revisão da parte computacional e da operacionalização destas ferramentas em ambiente SIG. Daqui decorre, o interesse em especificar os parâmetros utilizados na configuração das ferramentas, e nos passos necessários para serem executadas.

A ordem de abordagem dos métodos segue a mesma sequência exposta no enquadramento conceptual pensada em função dos objectivos de trabalho. O ponto de partida é validar a autocorrelação espacial e determinar se a distribuição dos dados é dispersa, concentrada ou aleatória. Quando avaliada a autocorrelação espacial, surge o interesse em perceber a evolução do grau de aglomeração (*clustering*), em função da variação da distância entre observações, e em comparar os padrões nos diferentes períodos, para avaliar quais as suas mudanças no tempo.

Apesar de, visualmente, se notar uma tendência para os visitantes se concentrarem nas áreas turísticas (como revisto na secção 5.2.), procura-se, com a aplicação destes primeiros métodos, verificar se o padrão é estatisticamente significativo e, como tal, suportado por um processo espacial subjacente. Sendo estas métricas derivadas das características globais da amostra, constatando-se a tendência comum das observações em se concentrarem, o que resta é identificar quais os locais de concentração (mais uma vez, considerando a sua significância estatística). Para o efeito, deve-se recorrer a métodos que consigam interpretar a heterogeneidade da autocorrelação espacial.

Assim sendo, optou-se pela ferramenta de análise de *cluster* e *outliers*, baseada no método *Local Moran Index*, para determinar quais as aglomerações estatisticamente significativas, sejam compostas por valores semelhantes elevados ou baixos, e quais os locais atípicos (locais onde o número de fotografias é diferente aos observados na sua vizinhança). O cálculo do índice implica estabelecer a noção de proximidade entre observações, sendo alcançada, neste caso, através da criação de uma matriz de pesos espaciais (ou matriz de proximidade espacial).

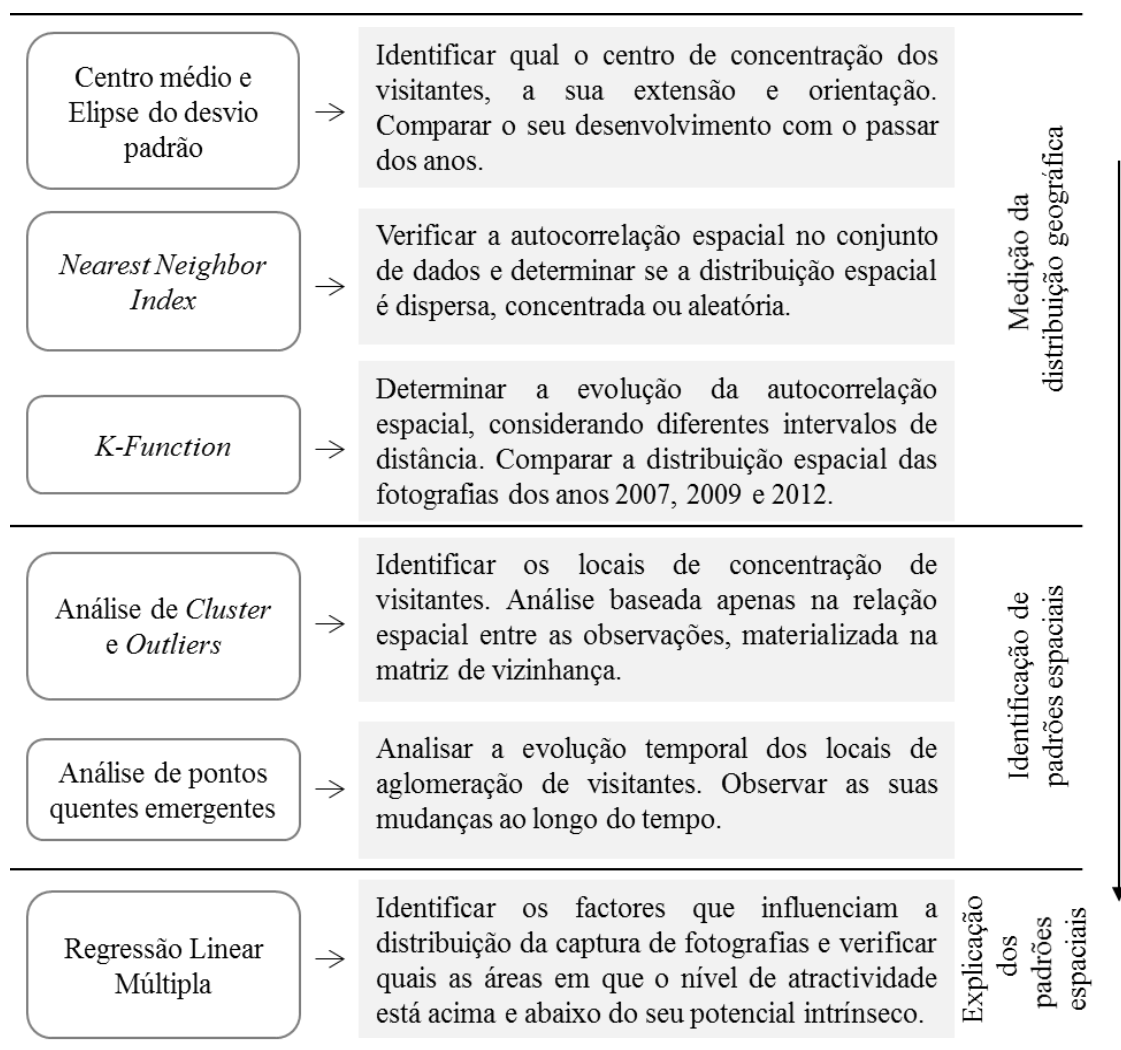
Todavia, importa discorrer qual a influência da componente temporal na concepção dos *clusters* espaciais. É neste sentido que, através da ferramenta *Emerging hot spot analysis*, se analisa, para além da relação espacial entre as fotografias, a sua indexação temporal face ao período em que foram carregadas. Assim, em complemento à localização dos focos, obtém-se como resultado o seu padrão sazonal (*i.e.* se os locais de concentração são consecutivos, esporádicos, etc.).

Previamente à identificação dos padrões espaciais, dedica-se uma parte da análise, à descrição das características da distribuição espacial das fotografias, com recurso a duas métricas, o centro médio e a elipse do desvio padrão. Por intermédio destas, tenta-se representar a compacidade e a orientação espacial do conjunto de dados. Com o intuito de avaliar se as características da distribuição diferem de um ano para outro, as medidas foram aplicadas anualmente de forma segmentada (para cada ano, desde 2007 a 2014).

Por fim, exploram-se as relações espaciais existentes entre o padrão observado dos visitantes e um conjunto de 24 variáveis seleccionadas que representam parte da oferta turística da cidade (*e.g.* Hotelaria, arquitectura religiosa, miradouros, etc. – Tabela 5.3). Para além disso, o que se pretende com a aplicação da análise de regressão linear múltipla,

é evidenciar se as áreas com potencial turístico são também as que concentram mais fotografias, *i.e.*, se são as mais cativantes para os visitantes.

Na Figura 5.19, apresenta-se o esquema processual da aplicação dos métodos estatísticos, sintetizando a sequência dos procedimentos e os propósitos da utilização dos métodos. Os detalhes para a operacionalização dos métodos são apresentados nas secções seguintes.



**Figura 5.19: Esquema processual da aplicação dos métodos de análise espacial.**

### 5.3.1. Centro médio e Elipse do desvio padrão

O cálculo do centro médio e da elipse do desvio padrão é simples, tanto em termos computacionais como na sua parametrização, sendo necessário apenas indicar qual a camada de informação com os dados a serem analisados, e especificar o *Case Field*, ou seja, o campo que identifica os grupos de dados. Quando se utiliza este parâmetro, as

observações são primeiro agrupadas conforme o identificador e, seguidamente, é calculada a estatística para cada grupo. Como se pretendia calcular as métricas para cada ano, neste parâmetro, seleccionou-se o campo “anos”, que discrimina as fotografias dos turistas segundo o ano em que foram carregadas (Anexo 13).

Na configuração da ferramenta elipse do desvio padrão, é preciso estabelecer os desvios padrões a considerar para o seu cálculo. Uma elipse calculada usando um desvio padrão, vai mostrar onde as observações estão mais concentradas mas, quando calculada, usando dois ou três desvios, vai mostrar onde a maioria dos eventos ocorre. Neste caso, as elipses foram apuradas considerando um desvio padrão. É de mencionar que, nesta análise, utilizaram-se as localizações originais das fotografias e não a sua agregação espacial.

### **5.3.2. *Nearest Neighbor Index***

Para o cálculo do índice do vizinho mais próximo, é necessário, para além dos dados de análise, indicar qual a extensão da área de estudo (uma vez que é utilizada para medir a distância média observada e esperada), e o critério de proximidade. A extensão da área obtém-se do *layer* que contém o limite da cidade. Em concordância com a unidade de medida pré-estabelecida no projecto, a área da cidade<sup>51</sup> foi apurada em m<sup>2</sup>, sendo igual a 86.877.691,93m<sup>2</sup>. No que diz respeito ao critério de proximidade, o ArcGIS tem disponíveis duas opções, as distâncias de Manhattan e Euclidiana, tendo-se optado pela segunda. Assim, a distância entre os vizinhos mais próximos será determinada a partir do segmento de recta que os une.

Nesta fase utilizaram-se os dados agregados em unidades hexagonais referidos na secção 5.1.3. Activando a opção *Generate report* do painel de controlo (Anexo 14), a ferramenta gera um relatório com informação bastante pertinente.

---

<sup>51</sup> A organização administrativa de Lisboa (Lei 56/2012 de 8 de Novembro, alterada pela Lei n.º 85/2015, de 7 de Agosto) criou um novo mapa da cidade. Esta acção consistiu na alteração geográfica das então 53 freguesias que passaram a ser 24, agregando algumas, delimitando outras de forma diferente e até criando uma nova área administrativa da cidade – o Parque das Nações. Esta última, para além do lugar com o mesmo nome, está situada entre o antigo limite norte do concelho de Lisboa e a foz do rio Trancão, território anteriormente pertencente às freguesias de Moscavide e Sacavém, do vizinho concelho de Loures. Esta alteração implicou o aumento da área oficial do concelho de Lisboa.



### 5.3.3. *K-Function (Multi-Distance Spatial Cluster Analysis)*

Como foi dito anteriormente, os dados a utilizar nesta análise correspondem às fotografias de três dos oito anos disponíveis, *i.e.*, 2007, 2009 e 2012. Estes anos foram escolhidos por apresentarem uma quantidade de fotos superior à da média anual, como se pode constatar na Figura 5.14. Antes de se executar a ferramenta, a amostra original de fotografias necessitou de ser segmentada por anos. Nesse sentido, sobre os dados (sem serem agregados) aplicou-se uma operação de selecção por atributos (*Select by attributes*), identificando apenas as fotografias enquadradas dentro dos três anos pretendidos, de forma independente para cada série. Após esta selecção, foi possível criar três novos temas com a informação das fotos para cada um dos três anos. Esta informação foi agregada em áreas hexagonais, seguindo o mesmo procedimento apresentado na secção 5.1.3. (a dimensão da célula da malha regular mantém-se igual) (ver Anexo 15).

Relativamente à parametrização (Anexo 16), a ferramenta permite a escolha do número de intervalos (de distância), para os quais a análise vai ser efectuada, sendo neste caso 20. Quanto maior o número, mais suave será a curva dos valores observados da função  $L(d)$ , permitindo identificar com mais facilidade a distância à qual o grau de *clustering* é maior. Para além disso, é possível estabelecer a distância de partida (onde se inicia a análise), e o valor de incremento para cada intervalo. Na distância de partida estabeleceu-se a distância mínima entre as observações e o seu primeiro vizinho mais próximo, calculada através da ferramenta *Calculate Distance Band from Neighbor Count*<sup>52</sup>. Desta forma, assegura-se que no início da análise, todas as observações tenham pelo menos um vizinho, garantindo assim a significância dos resultados. A distância resultante foi de 50m. O valor do incremento para os intervalos foi estabelecido, tomando como referência a distância média resultante da análise do índice do vizinho mais próximo, *i.e.*, a distância média entre as observações e os seus vizinhos mais próximos (65m).

No que concerne aos limites de confiança - altos e baixos - estes foram determinados com base num teste de 999 permutações, obtendo assim, um nível de confiança de 99,9% nos resultados (o nível máximo de significância estatística). De

---

<sup>52</sup> A ferramenta calcula as distâncias mínima, máxima e média considerando o *n*-ésimo vizinho mais próximo segundo o especificado, para um conjunto de *features*. A quantidade dos vizinhos para qual a análise vai ser executada é um parâmetro de entrada. Em *ArcGIS help*, consultado em Fevereiro, 2016. Url: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/calculate-distance-band-from-neighbor-count.htm>

referir que, ao escolher um teste de 999 permutações é exigido ao computador um grande esforço de cálculo que se traduz num incremento considerável nas horas de computação.

Como a análise foi realizada a partir dos dados agregados, para indicar o peso (a quantidade de fotografias) das unidades hexagonais, na opção *Weight field*, indicou-se o campo que continha o número de fotografias associadas a cada célula. Desta forma, para o cálculo dos valores de  $L(d)$  não será apenas considerada a distância entre as observações mas também a sua ponderação. Por fim, para que a análise se ajustasse à extensão da área de estudo, indicou-se no campo *Study area*, o *layer* com o limite da cidade.

Os resultados destas análises, para cada ano, são apresentados em gráficos onde se confrontam os valores observados e esperados da função  $L(d)$ , e os limites de confiança estatísticas.

#### **5.3.4. Análise de Clusters e Outliers**

O método da análise de *clusters* e *outliers* assenta na diferença entre os valores associados às entidades e na relação espacial entre elas. No cálculo do índice local de Moran, compara-se o valor (o número de fotos) de cada entidade ou *target features* com os valores dos seus vizinhos, sendo o processo realizado para cada unidade, continuamente, até analisar todas as observações.

A fim de executar a análise, torna-se necessário definir: i) a área circundante (área que abrange as entidades vizinhas a comparar) a cada *target feature*; e ii) a natureza da relação espacial entre as entidades. A relação espacial é definida com base na interacção entre observações, sendo, na maioria dos casos, representada através da distância euclidiana entre elas ou através da sua adjacência (entidades contíguas que partilham fronteiras) (Mitchell, 2005).

Como referido no enquadramento conceptual, no contexto dos SIG, uma vizinhança é implementada através dos pesos espaciais atribuídos às entidades “vizinhas” conforme o tipo de vizinhança definido (*e.g.* matriz de pesos espaciais ou matriz de proximidade).

No *software* utilizado, é possível criar uma matriz de pesos utilizando a ferramenta *Create Spatial Weight Matrix*, do módulo *Modeling Spatial Relationship Toolset*. No caso particular desta análise, a vizinhança foi baseada no declínio da função da distância (euclidiana). Deste modo, na parametrização da ferramenta (Anexo 17),

escolheu-se a opção “Inverse distance” como o método para conceptualizar a relação espacial entre observações. Assim, a influência das *features* vizinhas vai decrescendo, na medida em que a distância entre elas (*target feature* e *n* vizinhas) vai aumentando. Assumindo que nem todas as observações dentro da área de estudo estão relacionadas espacialmente, estabeleceu-se um limite de distância ou *threshold* para o raio de influência da vizinhança, igual a 135m (este valor foi obtido através da ferramenta *Incremental Spatial Autocorrelation*<sup>53</sup>, aplicada ao mesmo conjunto de dados).

Relativamente ao processo de cálculo, a ferramenta determina a relação entre pares de entidades e atribui um peso a cada par. No que se refere à vizinhança baseada na distância, o valor do peso é definido utilizando a própria distância observada entre as *features* (no caso das células hexagonais, a distância é medida a partir dos centroides das unidades). Os pesos espaciais são armazenados numa tabela, com uma linha e uma coluna para cada entidade, preenchidas com os valores correspondentes a cada par.

Como resultado, a ferramenta gera um ficheiro *SWM* que contém a informação relativa aos pesos. É este ficheiro que posteriormente será utilizado na análise de *clusters* e *outliers*, a fim de aquilatar qual a relação espacial entre as observações.

Para além da conceptualização da relação espacial entre entidades, na parametrização da ferramenta Análise de *Cluster* e *Outliers* (Anexo 20), só resta indicar o *layer* com as fotografias tiradas por visitantes, e o campo da sua tabela de atributos que contém os valores a ser analisados, *i.e.*, o número de fotos associadas a cada célula. Lembre-se que os dados a serem analisados são as fotografias dos turistas, agregadas por unidades hexagonais, referentes a todos os anos (2007 - 2014).

### **5.3.5. Análise de pontos quentes emergentes**

Para executar a análise de pontos quentes emergentes (APQE), deve-se primeiro estruturar os dados de entrada num cubo espaço-temporal, em formato NetCDF. Para o efeito, recorre-se à ferramenta *Create space time cube*. Na sua parametrização (Anexo

---

<sup>53</sup> A ferramenta calcula a autocorrelação espacial (o índice global de Moran) para uma série de distâncias. Como resultado, gera um gráfico que confronta os índices para os intervalos de distância com os seus respectivos *z-scores* (significância estatística). Os valores do *z-score* reflectem a intensidade de concentração espacial, e as subidas registradas no *score* identificam as distâncias onde o processo espacial que promove o agrupamento é mais pronunciado. A parametrização da ferramenta e os resultados são apresentados no Anexo 18 e Anexo 19. Em ArcGIS *help*, consultado em Fevereiro, 2016. Url: <http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/incremental-spatial-autocorrelation.htm>

21), identifica-se o *layer* referente às fotografias dos turistas no período 2007 – 2014, sem serem agregadas, e o campo (*data\_upload*) que contém a informação temporal. A seguir, indica-se qual o intervalo dos “passos de tempo”, ou seja, a dimensão temporal em que os dados vão ser agregados. Optou-se por uma agregação de um mês a fim de mitigar, ou mesmo omitir, o desvio/erro potencialmente presente na referência temporal dos dados (ver secção 2.3.2.1.).

O próximo parâmetro a ser definido tem a ver com o alinhamento temporal. Das três opções disponíveis, escolheu-se o alinhamento conforme a data de término do conjunto de dados. Assim, a agregação concentra-se sobre as observações mais recentes. A análise tem início no dia 31 de Dezembro de 2006 e termina 31 de Julho de 2014, com intervalos mensais. De referir que esta última data não corresponde ao registro mais recente (que é de 11 de Agosto de 2014), pois os últimos dias foram cortados de modo a ter intervalos de passos de tempo que coincidam com as divisões mensais.

Considerando que a agregação dos dados se baseia apenas na contagem das *features* envolvidas em cada caixa, os parâmetros da ferramenta relacionados com as sínteses estatísticas (*Summary fields*) são deixados em branco.

Como último parâmetro, define-se qual a dimensão espacial das caixas de agregação. Seguindo a mesma lógica da agregação espacial efectuada na análise de *clusters* e *outliers*, a dimensão de caixa escolhida corresponde à extensão média dos quarteirões da cidade,  $\approx 2.346\text{m}^2$ . No ArcGIS Pro 10.3. (ou na versão 10.4. desktop) também se pode escolher o tipo de geometria da caixa, continuando-se assim com uma célula hexagonal.

Como resultado, obtém-se um novo ficheiro de formato *.nc* que contém os dados agregados e um relatório do processo. No cubo agregaram-se 17.275 pontos (fotografias) em 65.192 caixas espaço-temporais, cobrindo a área de estudo e um período de 91 meses. O ficheiro *.nc* será utilizado na ferramenta de análise de pontos quentes emergentes.

Na parametrização da ferramenta APQE (Anexo 22), colocou-se especial atenção a definição da vizinhança espacial e temporal. No que respeita à primeira, determinaram-se como vizinhas, todas as caixas até uma distância de 135m da caixa *target* (toma-se como referencia o centroide da célula num contexto 2D). Como referido na parametrização da ferramenta de análise de *clusters* e *outliers*, com base nos resultados

do *Incremental spatial autocorrelation*, concluiu-se que esta é distância até onde o processo que promove o agrupamento é mais pronunciado.

Relativamente à segunda questão, optou-se por uma vizinhança temporal que considerasse um período anual para a mesma localização (caixa). Assim, indicou-se neste parâmetro o valor 12 (deve-se recordar que a análise inclui o intervalo temporal da caixa *target*).

O ArcGis Pro possibilita a visualização, num contexto 3D, do padrão das localizações, neste caso, considerando cada um dos intervalos mensais. Portanto, com o intuito de facilitar o entendimento da formação dos *hot spots* emergentes, apresenta-se como exemplo a visualização dos pontos localizados na zona centro da cidade (com ênfase na zona próxima à Praça do Comércio).

A visualização em 3D consegue-se através da ferramenta *Visualize Space Time Cube in 3D* (Anexo 23), sendo apenas necessário introduzir o ficheiro referente ao cubo de dados, identificar o campo que contém a contagem das observações “count”, e definir como tema de visualização, no parâmetro *Display theme*, a opção *Hot and cold spots results* (esta opção torna-se disponível só depois de ter corrido a ferramenta APQE). A representação em 3D é apresentada em conjunto com os resultados da APQE.

### **5.3.6. Regressão linear múltipla**

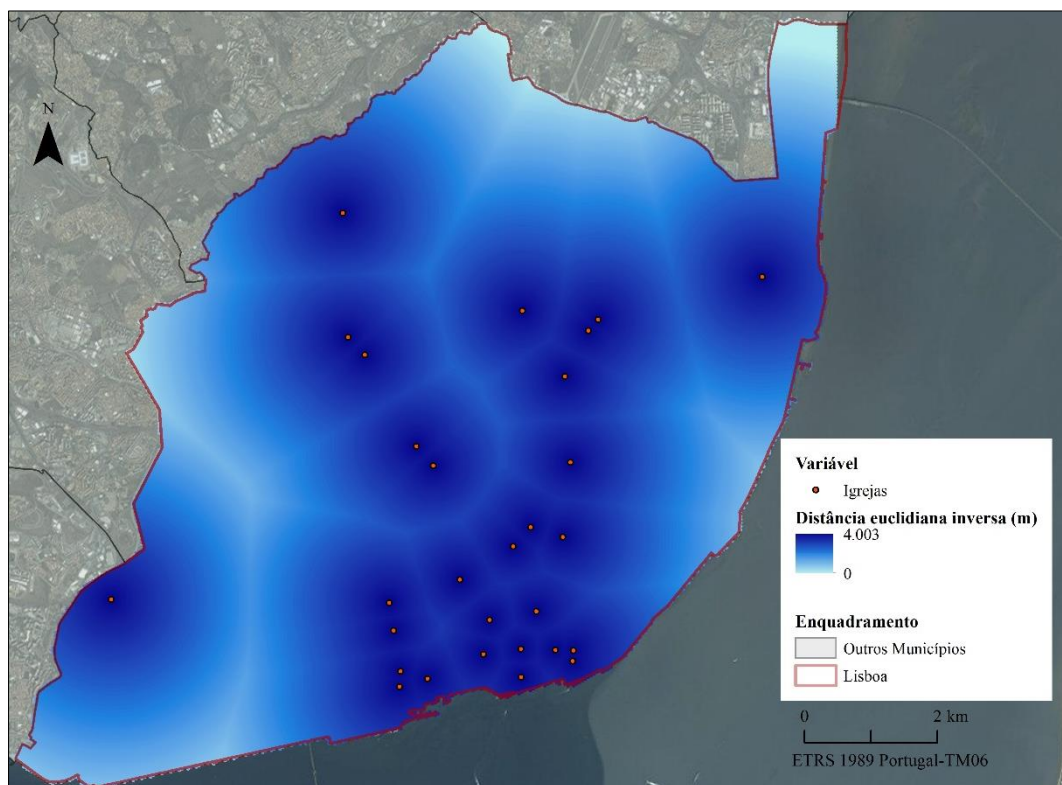
Um dos primeiros aspectos a serem considerados na esquematização de um modelo de regressão linear é a selecção dos factores explicativos (variáveis) a utilizar. Uma variável independente adicional pode melhorar a previsão da variável dependente.

De acordo com a informação disponibilizada pelo sítio internet da Câmara Municipal de Lisboa (CML)<sup>54</sup>, no separador referente às actividades turísticas, foram seleccionadas 24 variáveis para serem utilizadas como variáveis independentes (Tabela 5.3), *i.e.*, como factores explicativos da aglomeração geográfica das fotografias. Espacialmente, todas estas variáveis estavam representadas através de *layers* de pontos. No Anexo 24, como exemplo, pode-se observar a localização de três das variáveis seleccionadas: Hotelaria, Arquitectura premiada e Igrejas.

---

<sup>54</sup> Dados abertos, Câmara Municipal de Lisboa. Consultado em Julho, 2016. url: <http://dadosabertos.cm-lisboa.pt/dataset?tags=Geodados>

Todas as variáveis foram transformadas em escalas de razão, através da aplicação do inverso da distância euclidiana ao elemento analisado mais próximo. Como a unidade de medida é sempre a mesma (distância em metros) não se revelou necessário proceder à normalização dos valores. Para o efeito, utilizou-se a ferramenta *Euclidean distance*, tendo na sua parametrização sido estabelecida uma dimensão de célula de 25m<sup>2</sup> (Anexo 25). De modo a exemplificar a transformação das variáveis, apresenta-se o mapa da distância euclidiana inversa (Figura 5.20) correspondente a variável ‘igrejas’ (V14).



**Figura 5.20: Inverso da distância euclidiana à variável Igrejas.**

A melhoria do desempenho preditivo de um modelo de regressão linear não está só relacionada com a correlação de uma determinada variável com a variável dependente, mas também com a correlação dessa variável com as demais variáveis independentes do modelo. Deve-se pois, avaliar as possíveis dependências entre os regressores.

**Tabela 5.3: Variáveis explicativas pré-seleccionadas**

Código	Variável	Código	Variável
V1	Arquitectura Civil	V13	Hotelaria
V2	Arquitectura Industrial	V14	Igrejas
V3	Arquitectura Militar	V15	Mercados
V4	Arquitectura Nobre	V16	Miradouros
V5	Arquitectura Premiada	V17	Monumentos de Interesse Público
V6	Arquitectura Religiosa	V18	Monumentos Nacionais
V7	Cemitérios	V19	Museus
V8	Conjuntos de Interesse Público	V20	Parques Infantis
V9	Docas de Recreio e Marinas	V21	Parques Merendas
V10	Elevadores e Ascensores	V22	Parques Recreativos
V11	Feiras	V23	Parques Urbanos
V12	Geomonumentos	V24	Teatros

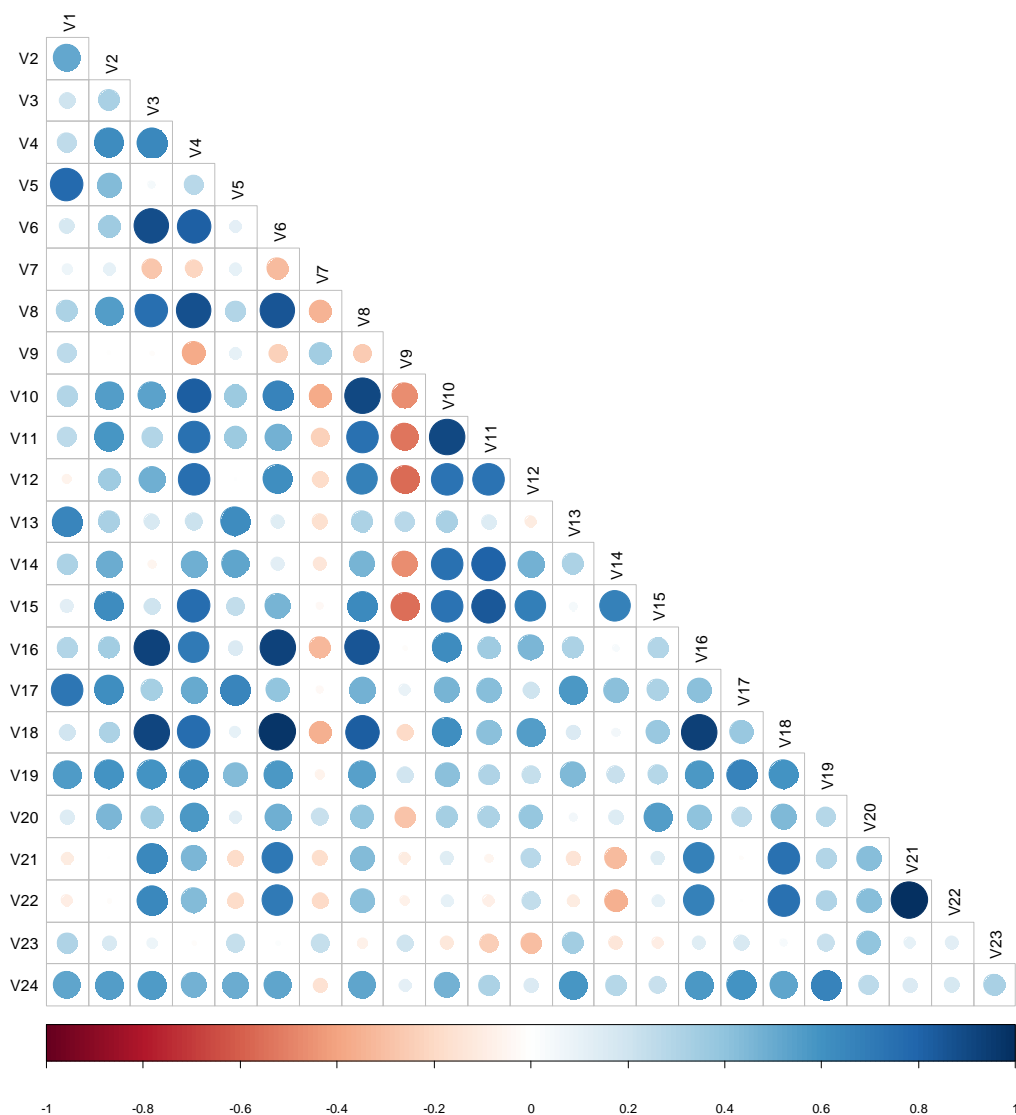
Em situações onde essas dependências forem fortes, pode-se afirmar que existe multicolinearidade. A multicolinearidade pode ser entendida como a existência de duas ou mais variáveis independentes numa análise de regressão múltipla (binária ou não) que se encontram altamente correlacionadas. A selecção e utilização das variáveis independentes que fazem parte do modelo deve ter em consideração a sua possível multicolinearidade.

O caso clássico de multicolinearidade ocorre quando nenhuma das variáveis explicativas na regressão é estatisticamente significativa. Desse modo, diz-se que duas ou mais variáveis explicativas são perfeitamente colineares se uma ou mais das variáveis puderem ser expressas como uma combinação linear da(s) outra(s) variável(eis).

Como a colinearidade significa que as variáveis independentes são correlacionadas, esta pode ser detectada, de entre outros modos, através da matriz de correlação entre as variáveis. Assim, de forma a prevenir a ocorrência de multicolinearidade entre variáveis foi calculada uma matriz de coeficientes de correlação de *Pearson* entre as variáveis independentes. Apenas as variáveis com correlações com valores de  $r > -0,85$  ou  $r < 0,85$  devem ser incluídas num modelo de regressão sem que o risco de multicolinearidade seja elevado.

Analisando o gráfico de regressão (Figura 5.21) pode-se observar que predominam as relações negativas fortes, enquanto as relações positivas são apenas moderadas. No primeiro caso, o que requer especial atenção, destaca-se a relação entre a variável V18 (Monumentos Nacionais) e as V3 (Arquitectura Militar), V5 (Arquitectura

Premiada) e V16 (Miradouros); a V16 com a V3 (Arquitectura Militar) e a V6 (Arquitectura Religiosa); a V11 (Feiras) com a V10 (Elevadores e Ascensores); a V10 com a V8 (Conjuntos de Interesse Público); a V8 com a V4 (Arquitectura Nobre) e a V6; e a V6 com a V3 (Arquitectura Militar).

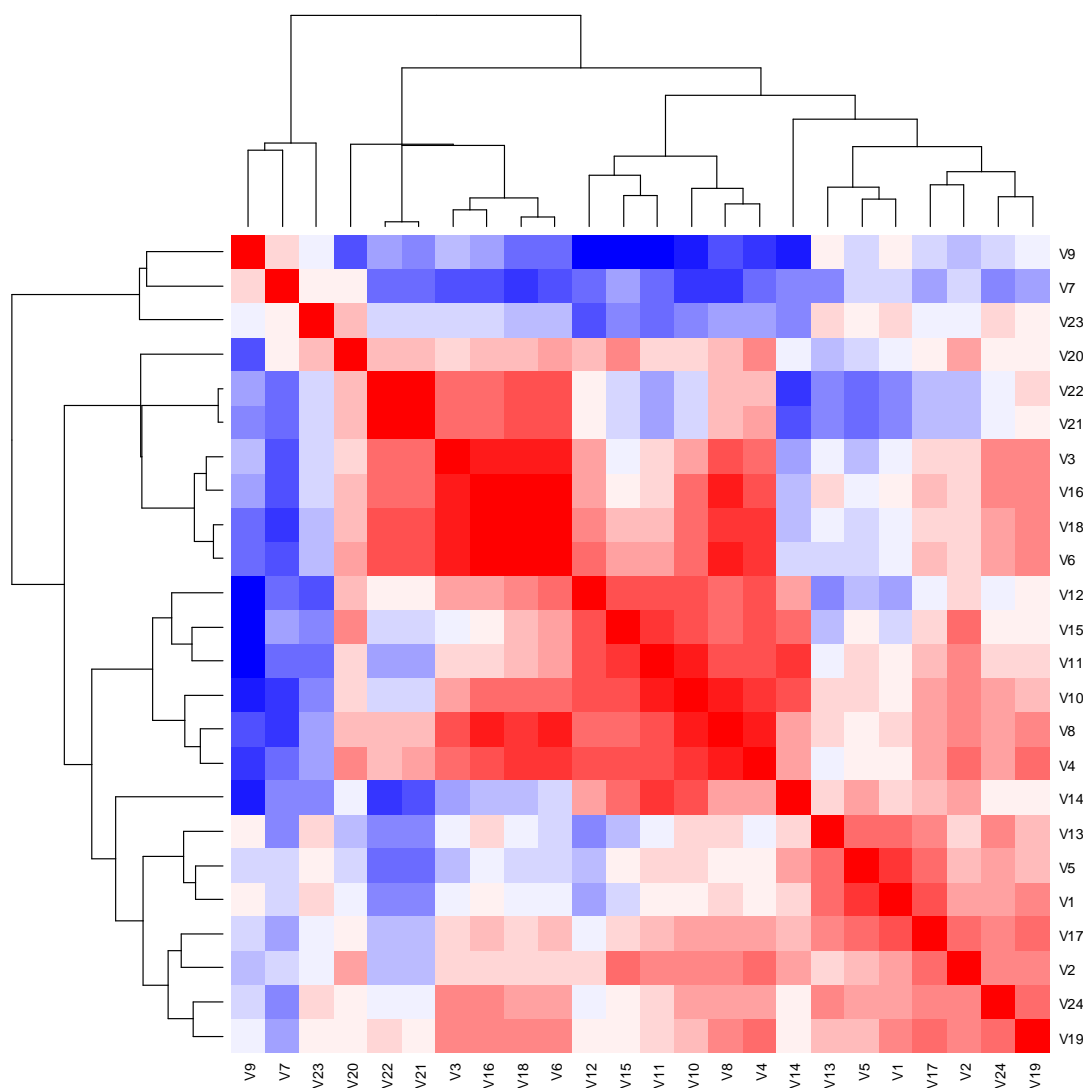


**Figura 5.21: Coeficiente de correlação de *Pearson* entre as variáveis pré-seleccionadas.**

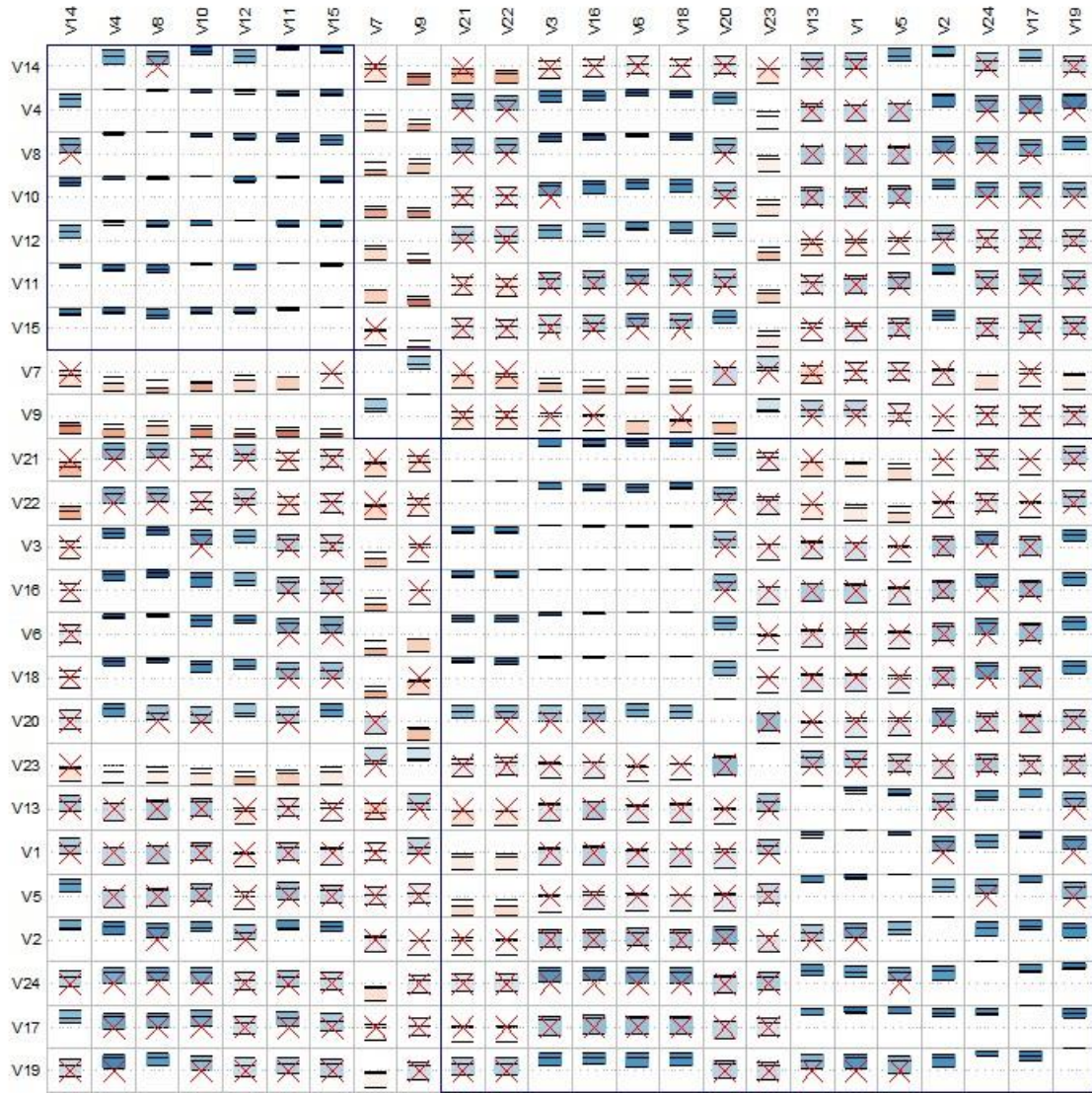
Observando o mapa de zonas quentes (*heat map*) representado na Figura 5.22, pode-se constatar que existem 3 *clusters*: um primeiro constituído pelas variáveis V3 (Arquitectura Militar), V6 (Arquitectura Religiosa), V16 (Miradouros), V18 (Monumentos Nacionais), V21 (Parques Merendas) e V22 (Parques Recreativos); outro de que fazem parte as variáveis V4 (Arquitectura Nobre), V8 (Conjuntos de Interesse Público), V10 (Elevadores e Ascensores), V11 (Feiras), V12 (Geomonumentos) e V15 (Mercados); e, por fim, o terceiro grupo, apresentando uma intensidade mais fraca, que



engloba as variáveis V1 (Arquitectura Civil), V2 (Arquitectura Industrial), V5 (Arquitectura Premiada), V13 (Hotelaria), V14 (Igrejas), V17 (Monumentos de Interesse Público), V19 (Museus) e V24 (Teatros). Esta fraqueza é perceptível na Figura 5.23, onde se pode constatar que o referido grupo (quadrante inferior direito) é quase todo composto por relações estatisticamente não significativas para um grau de confiança de 95%.



**Figura 5.22: Clusters de variáveis e zonas quentes.**



**Figura 5.23: Significância estatística das correlações e dos clusters.**

A presença de multicolinearidade pode ser detectada de várias maneiras. Outras técnicas muito usadas são o valor de tolerância ou o seu inverso, designado por factor de inflação de variação (*Variance Inflation Factor*), cujos valores elevados indicam a existência de colinearidade.

$$VIF = \frac{1}{1 - R_j^2} \quad (5.1)$$

Sendo que  $R_j$  resulta da regressão de  $x_j$  com as outras variáveis, *i.e.* o coeficiente de determinação da equação de regressão no passo 1. Sugere-se, no entanto, quando ocorrer colinearidade, que algumas variáveis explicativas sejam retiradas do estudo e/ou que se tente obter um maior número de observações.

O *VIF* é uma medida do grau em que cada variável independente é explicada pelas demais variáveis independentes. Quanto maior for o factor de inflação da variância, mais severa será a multicolinearidade. Efectivamente, o *VIF* mede a redundância entre as variáveis explicativas. Como regra geral, as variáveis explicativas com valores de *VIF* superiores a 7,5 devem ser removidas (uma-a-uma) do modelo de regressão. Na Tabela 5.4 está disponibilizada a contabilização dos valores de *VIF* para cada uma das 24 variáveis pré-seleccionadas. Destas, 12 foram excluídas por apresentarem valores de *VIF* superiores a 7,5.

**Tabela 5.4: Valores de VIF.**

<b>Código</b>	<b>Variável</b>	<b>VIF</b>
<b>V1</b>	Arquitectura Civil	5,390697
<b>V2</b>	Arquitectura Industrial	5,870192
<b>V3</b>	Arquitectura Militar	<b>20,850831</b>
<b>V4</b>	Arquitectura Nobre	<b>16,825982</b>
<b>V5</b>	Arquitectura Premiada	4,50322
<b>V6</b>	Arquitectura Religiosa	<b>45,079639</b>
<b>V7</b>	Cemitérios	3,499951
<b>V8</b>	Conjuntos de Interesse Público	<b>104,929136</b>
<b>V9</b>	Docas de Recreio e Marinas	6,700964
<b>V10</b>	Elevadores e Ascensores	<b>114,605176</b>
<b>V11</b>	Feiras	<b>22,373294</b>
<b>V12</b>	Geomonumentos	5,699458
<b>V13</b>	Hotelaria	6,158302
<b>V14</b>	Igrejas	<b>16,869859</b>
<b>V15</b>	Mercados	<b>10,306751</b>
<b>V16</b>	Miradouros	<b>58,239621</b>
<b>V17</b>	Monumentos de Interesse Público	3,962245
<b>V18</b>	Monumentos Nacionais	<b>58,774193</b>
<b>V19</b>	Museus	6,248456
<b>V20</b>	Parques Infantis	4,139325
<b>V21</b>	Parques Merendas	<b>359,060052</b>
<b>V22</b>	Parques Recreativos	<b>442,869192</b>
<b>V23</b>	Parques Urbanos	2,993635
<b>V24</b>	Teatros	4,270818

Posto isto, foi realizada uma análise exploratória em que foram corridos modelos que permitissem a existência de 1 a 5 variáveis independentes. Foram testadas 35.526 combinações das quais 6.866 apresentaram um  $R^2$  ajustado superior a 0,9. Dentro deste grupo mais restrito, os 38 modelos com melhor desempenho recorreram sempre a 5

variáveis, com desempenhos preditivos entre 0,95 e 0,96. Na Tabela 5.5 identificam-se as variáveis que foram seleccionadas por esses 38 modelos e a frequência com que o foram.

**Tabela 5.5: Variáveis referidas e respectivas frequências.**

Código	Variável	Frequência (Nº)
V13	Hotelaria	33
V14	Igrejas	32
V10	Elevadores e Ascensores	29
V12	Geomonumentos	21
V15	Mercados	8
V9	Docas de Recreio e Marinas	6
V21	Parques Merendas	5
V22	Parques Recreativos	5
V20	Parques Infantis	3
V16	Miradouros	2
V4	Arquitectura Nobre	1
V5	Arquitectura Premiada	1
V17	Monumentos de Interesse Público	1
V24	Teatros	1

Do cruzamento da Tabela 5.4 com a Tabela 5.5, obteve-se o conjunto final de 7 variáveis independentes a utilizar no modelo (Tabela 5.6).

**Tabela 5.6: Variáveis seleccionadas para o modelo.**

Código	Variável	Pesos
V13	Hotelaria	0,10
V12	Geomonumentos	0,12
V9	Docas de Recreio e Marinas	0,21
V20	Parques Infantis	0,03
V5	Arquitectura Premiada	0,15
V17	Monumentos de Interesse Público	0,29
V24	Teatros	0,10

Como medidas de desempenho do modelo utilizaram-se os valores de  $R^2$  e  $R^2$  ajustado. O seu intervalo de valores possíveis oscila entre 0,0 e 1,0. Na prática não é expectável encontrar uma correlação perfeita ( $R^2=1$ ) pois existem muitos fatores que influenciam o comportamento de uma variável. Assim, a qualidade da regressão pode ser dada por:

$$R^2 = \frac{V_e}{V_t} \quad (5.2)$$

Onde  $R^2$  é o coeficiente de determinação e,  $V_e$   $V_t$  são, respectivamente, a variação explicada e total de Y.

O valor de  $R^2$  ajustado é sempre um pouco menor do que o valor de  $R^2$ , porque reflecte a complexidade do modelo (*i.e.*, o número de variáveis) e como este se relaciona com os dados. Consequentemente, constitui uma medida mais exacta do desempenho do modelo. Adicionar uma variável explicativa ao modelo provavelmente aumentará o valor de  $R^2$ , mas pode diminuir o valor de  $R^2$  ajustado. Suponha-se que se pretende criar um modelo de regressão de fotografias (o número de fotografias associadas a cada subsecção estatística é a variável dependente, y). Assim, o  $R^2$  ajustado é dado por:

$$R_a^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k} \quad (5.3)$$

Onde  $R_a^2$  é o  $R^2$  ajustado,  $n$  a dimensão da amostra e  $k$  o número de variáveis independentes.

Um valor de  $R^2$  ajustado de 0,68 (Tabela 5.7) indica que o modelo (as variáveis explicativas modeladas recorrendo a uma regressão linear) explica cerca de 68 por cento da variação ocorrida na variável dependente. Dito de outro modo, o modelo traduz cerca de 68 por cento do histórico de fotografias tiradas.

**Tabela 5.7: Desempenho do modelo.**

Número de observações	17.604
R quadrado	0,682851
R quadrado ajustado	0,682725



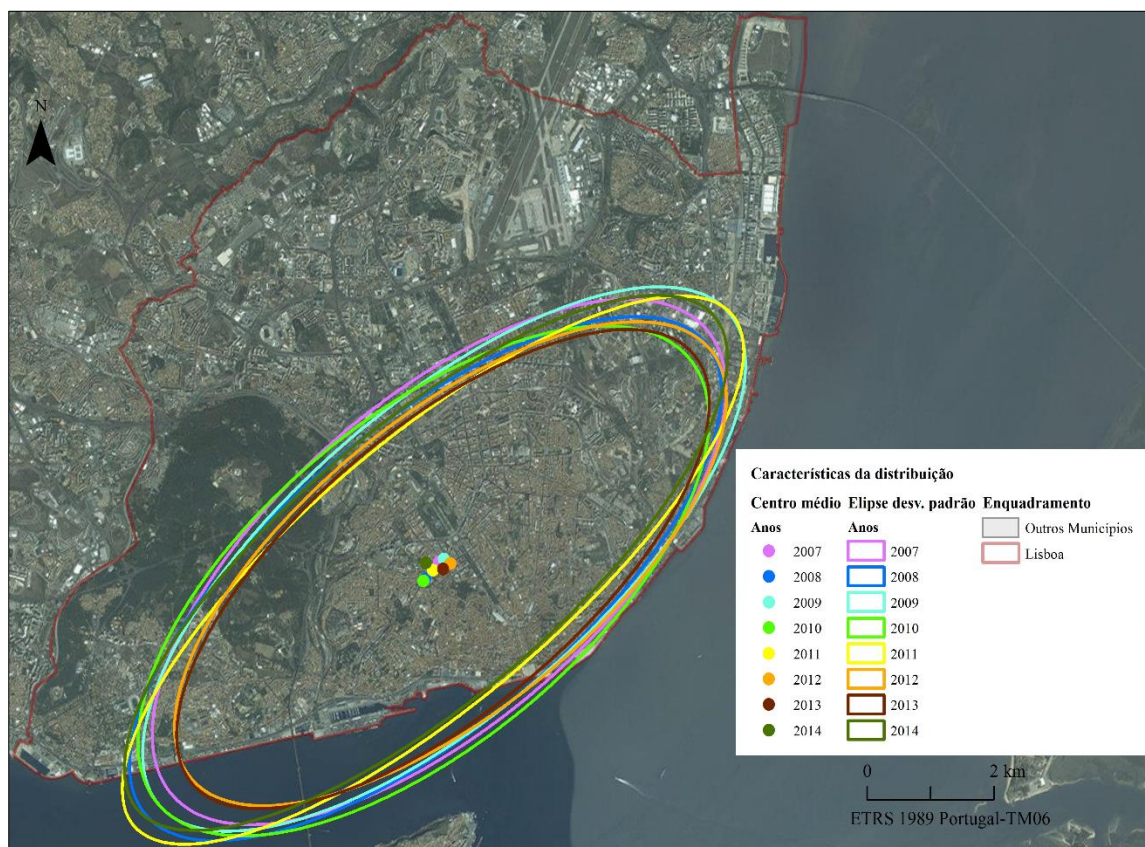


## 6. Análise dos resultados

Nesta secção são discutidos os resultados obtidos pela aplicação dos métodos estatísticos apresentados nos capítulos anteriores. Para uma melhor compreensão, estrutura-se a análise dos resultados seguindo a mesma sequência referida na secção 5.3.

### 6.1. Centro médio e Elipse do desvio padrão

Segundo o cálculo da medida central (Figura 6.1), identifica-se o sector em torno da Av. Braamcamp (no extremo noroeste da freguesia Santo António), entre o Marquês de Pombal e o Rato, como o mais central. O sector identificado apenas serve como referência e não deve ser entendido como o local para onde convergem os turistas. A localização dos centros médios sobre este sector evidencia a sua tendência de se afastarem do interior, situando-se mais para o quadrante sul - sudoeste da cidade.



**Figura 6.1: Características da distribuição, Centro médio e Elipse do desvio padrão.**

Observando com maior detalhe (Anexo 26), a localização dos centros médios não varia significativamente (de todos os centros, os mais afastados [2010 - 2012] distam entre si 498m em linha recta), mantendo-se sempre em torno desse sector. Consta-se assim, uma pequena mudança na localização dos centros para Sudoeste o que reflecte a orientação da distribuição espacial da amostra, influenciada pelas fotografias situadas na parte Sul do eixo ribeirinho da cidade.

À semelhança dos resultados obtidos no cálculo dos centros médios, as elipses do desvio padrão também são coincidentes. Em todos os anos, as elipses apresentam uma maior cobertura do litoral do que do interior, enfatizando a importância do eixo ribeirinho e das atracções situadas nas suas proximidades (*e.g.* Praça de Comércio, Mosteiro dos Jerónimos, Padrão dos Descobrimentos, entre outros).

As elipses expandem-se mais para o sudoeste do que para o nordeste, chegando a envolver, nessa direcção, a área circunvizinha ao Padrão dos Descobrimentos. De referir que nenhuma das elipses consegue abarcar, no lado oposto, os pontos de interesse do Parque das Nações, o que demonstra a disparidade existente quanto ao número de observações localizadas em ambos os extremos. Tanto no eixo maior (no sentido nordeste - sudoeste) como no eixo menor (no sentido interior - litoral), as elipses evidenciam uma sobreposição quase perfeita, sem mudanças pronunciadas no que respeita às coberturas.

A diferença mais marcante surge quando se comparam as elipses dos anos de 2011 e 2013, sendo a primeira mais dilatada do que a segunda (*i.e.* maior eixo principal). Centrando as atenções apenas no eixo menor, as diferenças não são tão evidentes e, neste caso, é a elipse do ano de 2010 a que apresenta um maior alargamento do litoral para o interior.

Em termos gerais, o que as medidas revelam é que a distribuição espacial dos visitantes ao longo dos 8 anos tem características muito semelhantes, denotando-se uma forte influência dos locais de interesse adjacentes ao litoral e das atracções situadas na zona de Belém.

## **6.2. Índice do vizinho mais próximo (NNI)**

Apesar de ser possível discernir visualmente que as fotografias dos visitantes tendem a concentrar-se em áreas específicas da cidade (*e.g.*, nas áreas descritas na secção 5.2.1.), é necessário constatar se esse padrão é estatisticamente significativo.



O resultado do NNI (Tabela 6.1 e Anexo 27) confirma que a distribuição espacial das observações é concentrada (sendo um padrão estatisticamente significativo, com  $z$ -score inferior a -2,56 e um  $p$ -value igual a 0) e que, como tal, não é produto de um processo espacial aleatório.

**Tabela 6.1: Resultados índice do vizinho mais próximo.**

Índice do vizinho mais próximo (NNI)	
Distância média observada	65,66m
Distância média esperada	71,75m
Índice do vizinho mais próximo	0,92
$z$ -score	-10,55
$p$ -value	0,00000

### 6.3. *K-function*

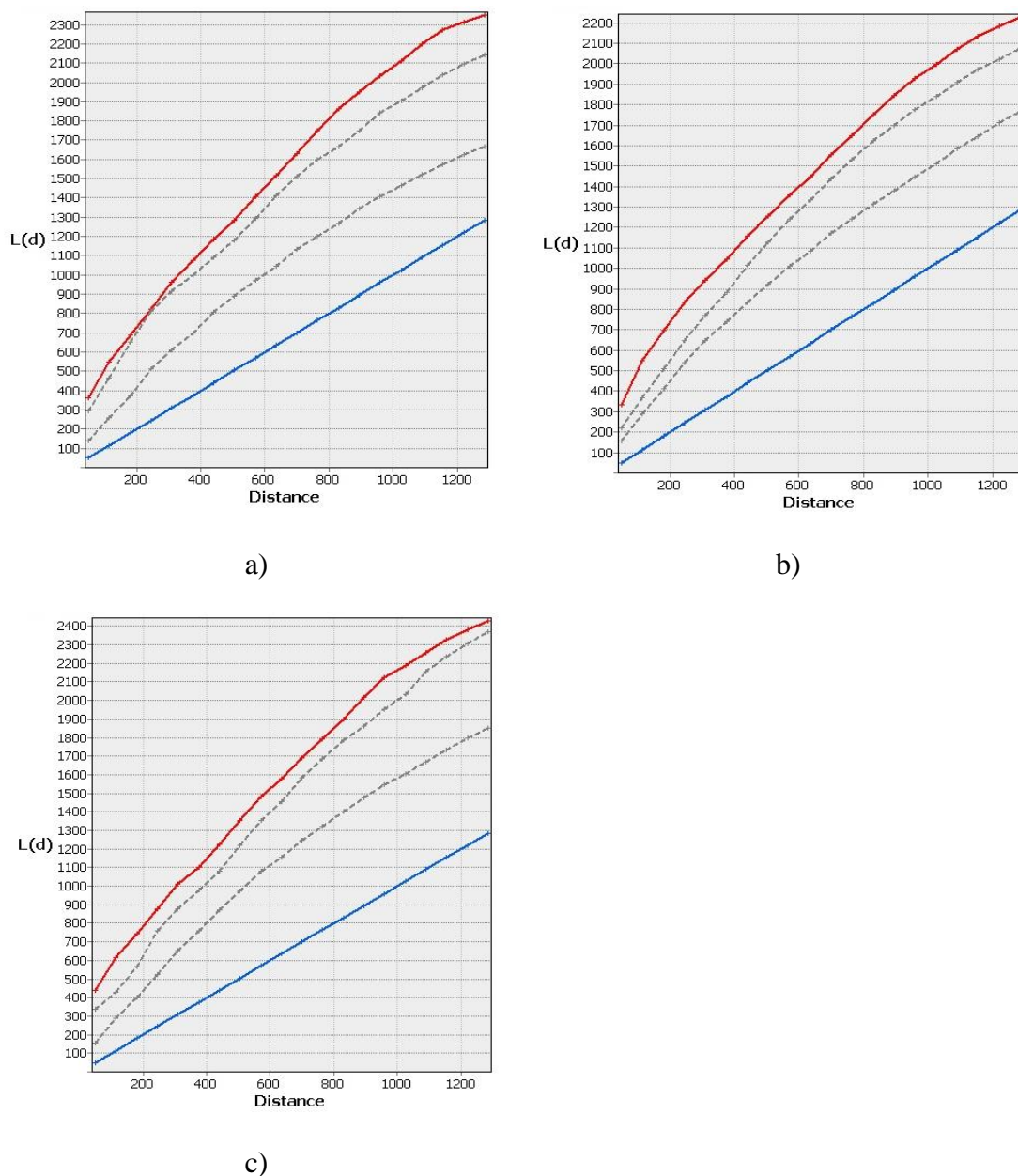
Nos gráficos da função  $L(d)$ , para os anos 2007, 2009 e 2012 (Figura 6.2), é visível que a distribuição das observações é constantemente concentrada ao longo dos intervalos de distância estabelecidos (dos 50m até os 1.300m). Em todas as séries de distâncias, o padrão de *clustering* é estatisticamente significativo (nos primeiros 400m a significância é mais acentuada, para os anos 2009 e 2012).

As linhas dos valores  $L(d)$ , nos três casos, tendem a manter-se em constante crescimento até aos 1.150m aproximadamente (ver Anexo 28) onde se pode então observar um decréscimo, indicando que a esta distância a significância estatística do padrão já é menos forte.

Em todos os anos, a função  $L(d)$  (observada) apresenta valores elevados quando comparados com os valores esperados, incluindo a distância de partida,. Estes valores vão aumentando à medida que a distância vai sendo incrementada. Isto significa que mais observações (fotografias) são encontradas à medida que o limiar da distância de análise se dilata. Neste sentido, o primeiro pico evidencia a presença de *clusters* a essa distância e, um segundo ou terceiro pico, podem indicar que os *clusters* antes evidenciados estão concentrados.

Os resultados confirmam que a distribuição espacial das observações é concentrada e que o padrão é estatisticamente significativo. Como se pode observar nos gráficos, a tendência de *clustering* é significativa, mesmo nas distâncias mais curtas, o que dá uma ideia sobre a estrutura da distribuição das fotografias. Em síntese, pode-se

concluir que, nos períodos de análise, as fotografias dos visitantes tendem a se concentrar em áreas de pequena dimensão e que essas áreas estão próximas umas das outras.



**Figura 6.2:** Função  $L(d)$  para as fotos agregadas dos anos a) 2007, b) 2009 e c) 2012.

#### 6.4. Análise de *Clusters* e *Outliers*

Como era de esperar, os *clusters* estão localizados nas zonas turísticas da cidade, e na circunscrição das atracções mais relevantes. Na secção 5.2.1 quando se apurou o número absoluto de fotografias associadas aos sítios de interesse turístico, foram as atracções conhecidas da cidade, as que mais se diferenciaram. Contudo, esta tentativa de

exploração não incluía aspectos importantes na análise, como a configuração da vizinhança dos locais ou a significância estatística dos dados apurados.

Olhando para o mapa (Figura 6.3), o facto mais saliente tem a ver com que a quase totalidade dos *clusters* espaciais são de categoria HH (*High-High*, segundo a nomenclatura imposta pela ferramenta), ou seja, são locais onde valores elevados estão rodeados por valores semelhantes (elevados). Portanto, tem-se uma grande afluência de visitantes, tanto nesses locais como nas suas áreas contíguas.



**Figura 6.3: Clusters e outliers espaciais do turismo em Lisboa.**

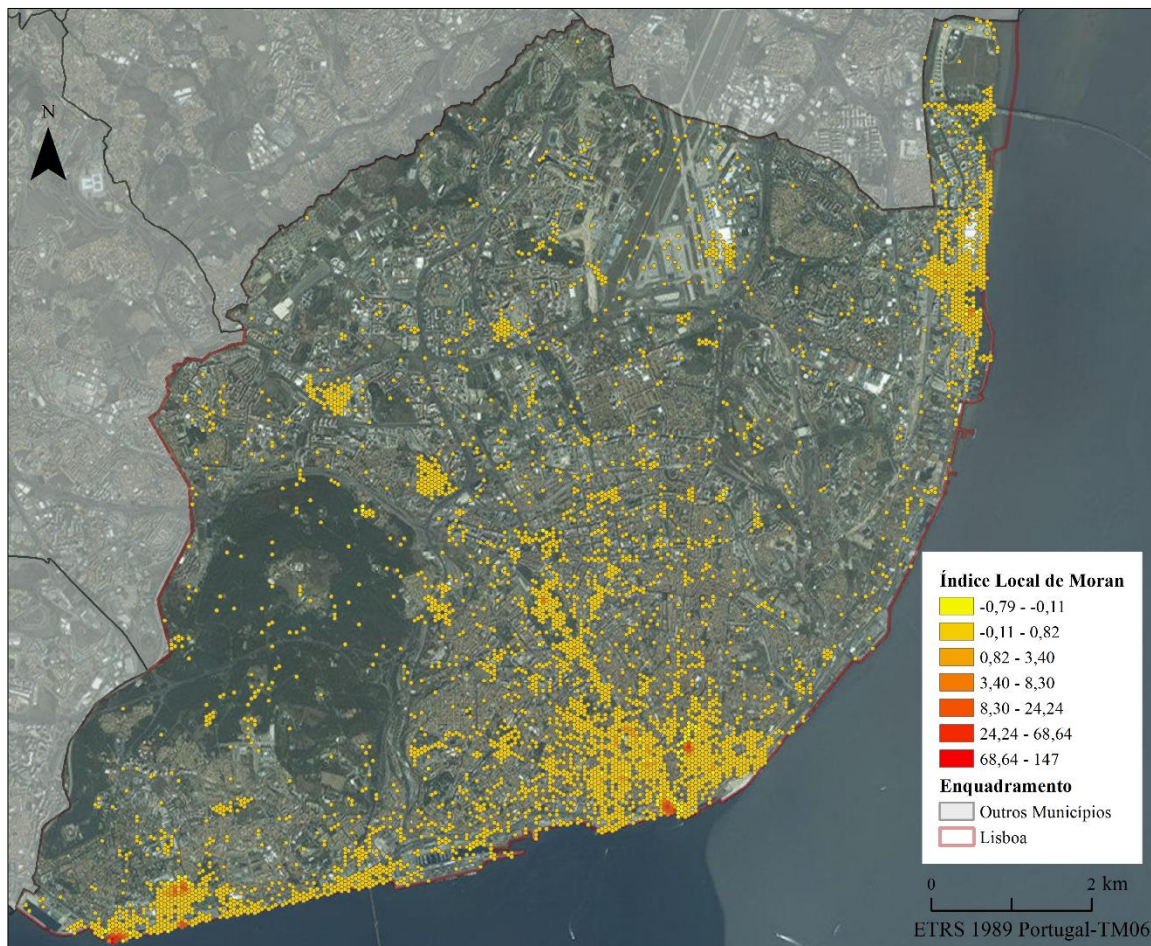
Por outro lado, apenas um *cluster* é de categoria diferente, LH (*Low-High*), e refere-se a um local que possui um valor baixo mas rodeado por valores elevados. Assim, este local particular é menos frequentado do que as áreas ao seu redor. Não existem *clusters* de tipo LL (*Low-Low*) ou *outliers* de tipo HL (*High-Low*).

Com efeito, verifica-se que as atracções turísticas funcionam como foco dos *clusters* espaciais. É sobre elas que se formam os *clusters*, e é a partir delas que estes se



expandem para as áreas próximas. Em muitos casos, a extensão dos *clusters* segue a forma geométrica das atracções (*e.g.* praças), não se expandindo além do perímetro delas.

Com o intuito de perceber qual a diferença entre os *clusters* espaciais no que respeita à sua magnitude, na Figura 6.4, são apresentados os valores do Índice de Moran [ $I_i$ ]<sup>55</sup> (a leitura deste pode ser complementada com o mapa que mostra os valores das células classificados por classes naturais, Anexo 29).



**Figura 6.4: Índice local de Moran.**

Os valores mais altos do  $I_i$  estão associados aos hexágonos situados sobre ou na adjacência das atracções (*e.g.* os monumentos da Torre de Belém, Padrão dos Descobrimentos e Marquês de Pombal). No caso do Mosteiro dos Jerónimos, e do Oceanário de Lisboa, os valores mais diferenciados correspondem às células situadas na parte da fachada, na área junto às entradas.

<sup>55</sup> A magnitude do  $I_i$  (alta ou baixa) depende: da diferença dos valores das *features*, dos vizinhos com valores semelhantes, e da magnitude dos valores do atributo. Quando as *features* vizinhas possuem valores semelhantes ou mesmo iguais à *target feature*, o índice resultará num valor elevado, ao contrário, quanto maior a diferença entre eles, mais baixo será o índice.

Relativamente aos *clusters* sobre as praças (*e.g.* Praça de Restaurados, Dom Pedro IV, de Comércio, Luís de Camões), as células com os valores mais salientes de  $I_l$  tendem a se situar mais perto do centro. Nas praças dos Restauradores e Dom Pedro IV, por exemplo, são as células centrais as que mais se destacam.

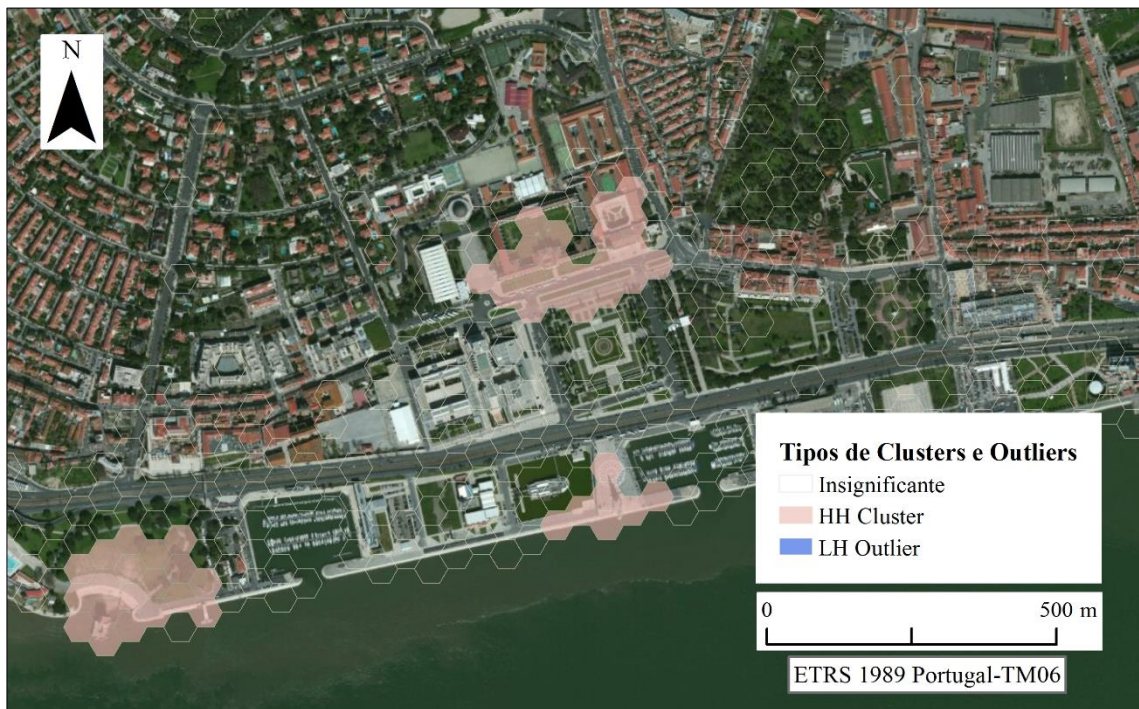
Com o apoio da leitura cruzada de ambos os mapas (Figura 6.3 e Figura 6.4), continua-se com a descrição da localização dos *clusters* e *outliers*, e com a identificação das células de maior destaque (com  $I_l$  elevados) de cada aglomerado. Para ajudar à interpretação dos mapas, na Figura 6.5, Figura 6.6 e Figura 6.7, expõem-se os locais de concentração a uma escala maior.

Os *clusters* estão localizados em três grandes zonas, no Sul, na freguesia de Belém, no Centro, nas freguesias de Santa Maria Maior e Santo António e, no Norte, no Parque das Nações.

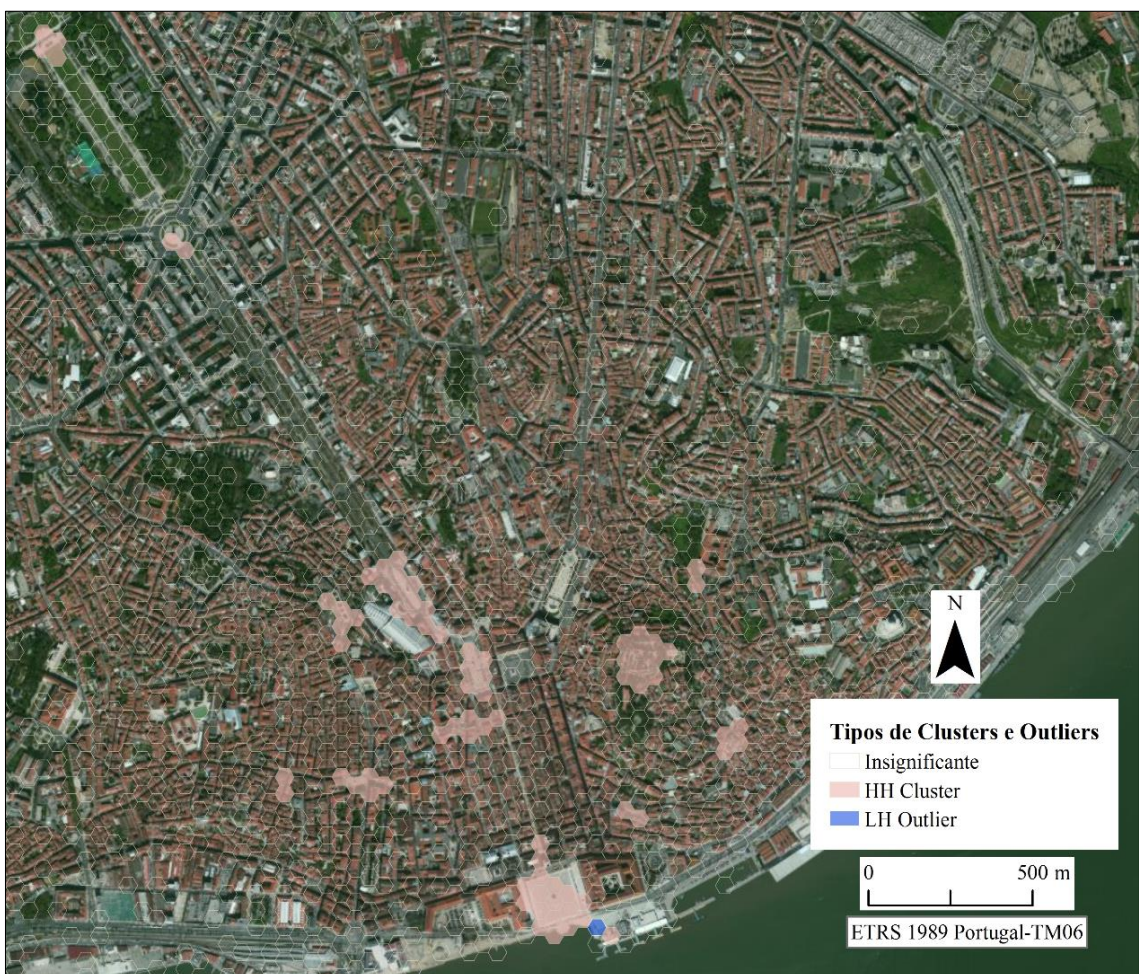
Na parte sul, existem três *clusters* sobre monumentos, o da Torre de Belém, o do Padrão dos Descobrimentos e o do Mosteiro dos Jerónimos (Figura 6.5 e Anexo 30). Dos três, o aglomerado situado sobre o Mosteiro é o de maior extensão da cidade (compreende 21 células) mas não o que possui o valor mais elevado do  $I_l$ . A mancha contorna o edifício, ainda incluindo parte da frente norte da Praça do Imperio e do jardim de Belém. É o *cluster* associado à Torre de Belém o que apresenta o valor mais elevado do  $I_l$ , da cidade e obviamente deste trio ( $I_{lmax}=146,40$  para a célula sobre o monumento). A extensão do aglomerado cobre quase todo o jardim da Torre, mais distendido para o lado nordeste do monumento. No Padrão dos Descobrimentos encontram-se os  $I_l$  menos salientes desta zona, e é também o grupo de menor extensão, cobrindo apenas a praça imediatamente subjacente ao monumento.

No centro, identificam-se 13 grupos de *clusters* (Figura 6.6 e Anexo 31), de cima para baixo, no Parque Eduardo VII, no Marquês de Pombal, nas praças dos Restauradores e Dom Pedro IV, no miradouro de São Pedro de Alcântara (incluindo a estação norte do Elevador da Glória), no Elevador de Santa Justa e Convento do Carmo, na praça Luís de Camões (incluindo o Largo de Chiado), na estação norte do Elevador da Bica, no Convento da Graça (incluindo o miradouro da Graça), no Castelo de São Jorge, nos Miradouros das Portas do Sol e de Santa Catarina, na Igreja Sé de Lisboa e na praça do Comércio (incluindo parte rua Augusta).





**Figura 6.5: Tipos de *clusters*, zona de Belém.**



**Figura 6.6: Tipos de *clusters* e *outliers*, zona centro.**

Nessa mesma ordem, os dois primeiros *clusters* apresentam uma forma estreita. No parque Eduardo VII, o aglomerado abarca o miradouro (do lado noroeste do parque), parte do Jardim Amália Rodrigues e uma porção do lado norte do eixo central do parque. No segundo, os *clusters* situam-se apenas na frente do monumento. Em ambos os casos, os valores do  $I_l$  das suas células não mostram variações pronunciadas.

Nas praças dos Restauradores e Dom Pedro IV, como já referido, os *clusters* tendem a ajustar-se às suas formas geométricas. Na primeira, a mancha estende-se um pouco para o sul até chegar a Estação do Rossio. No miradouro de São Pedro de Alcântara, o aglomerado abrange a estação norte do Elevador da Glória e parte do percurso. Neste caso, o valor de  $I_l$  (2,08) mais elevado está relacionado a célula que situada do lado da varanda (do miradouro).

Outro aglomerado tem como foco o Elevador de Santa Justa (a célula sobreposta tem um  $I_l$  [2,17 e  $n$  fotos = 46] mais diferenciado do que os restantes). Este apresenta uma forma alongada, acompanhando a rua de Santa Justa em direcção ao Largo do Carmo.

A praça Luís de Camões e o Largo do Chiado formam um outro aglomerado. A sua mancha expande-se em sentido Este-Oeste, desde a entrada da estação do metro até o início da rua do Loreto. A semelhança do que acontece com as outras praças, a célula que se sobrepõe à estátua de Luís de Camões é a de índice mais elevado ( $I_l = 2,56$  e  $n$  fotos = 60). Muito perto, encontra-se o *cluster* do Elevador da Bica, de extensão restrita, que encerra a área da estação (norte) e parte do percurso do próprio elevador.

Em Alfama, localizam-se quatro grupos de *clusters*. O maior deles com uma abrangência espacial que corresponde ao Castelo de São Jorge. Como já foi dito, a célula que dá para a porta de ingresso constitui o foco ( $I_l=30$ ). Os restantes *clusters*, o Convento da Graça, os Miradouros das Portas do Sol e de Santa Catarina, e a Igreja Sé de Lisboa, são aglomerados de reduzidas dimensões. Na igreja, as duas únicas células que se destacam possuem valores semelhantes. No convento e nos miradouros, as células mais relevantes são a da fachada, no primeiro caso, e a do terraço, no segundo caso.

O último *cluster* desta zona localiza-se sobre a praça do Comércio. É o aglomerado de maior dimensão deste grupo e estende-se desde o parque de estacionamento da estação fluvial, passando pelo Cais das Colunas, até à rua Augusta (chega à rua de São Julião). Aqui, está localizada a célula com o maior número de fotografias ( $n = 374$ ). Neste caso, o foco do aglomerado é a célula sobre a estátua ( $I_l=36$ ),



contudo, devem-se também ressaltar as células próximas ao rio que detêm índices igualmente significativos.

Na parte norte, identificam-se três *clusters* (Figura 6.7 e Anexo 32), dentro do sector do Parque das Nações. O mais extenso, prolonga-se desde o edifício do Oceanário, em direcção a Sul, passando pela esplanada Dom Carlos I, até o Jardim da Água. A célula virada para a fachada é a que apresenta o valor máximo do  $I_l$  (8,30) e de fotos (195). As demais células que conformam o aglomerado estão posicionadas à volta do Oceanário, seguindo a forma dos passeios do complexo do Parque das Nações.

Os dois *clusters* mais para Norte situam-se, um, do lado Este do Centro Comercial Vasco da Gama, na rua Alameda dos Oceanos (frente ao Rossio dos Olivais) e o outro, entre o centro comercial e a Estação de metro Oriente. O primeiro *cluster* tem como foco a estátua monumental ( $I_l = 1,89$ ;  $n$  fotos = 62) e, no segundo, a célula mais diferenciada é a que envolve a área do passeio que conecta ambos os elementos ( $I_l = 2,80$ ;  $n$  fotos = 68).

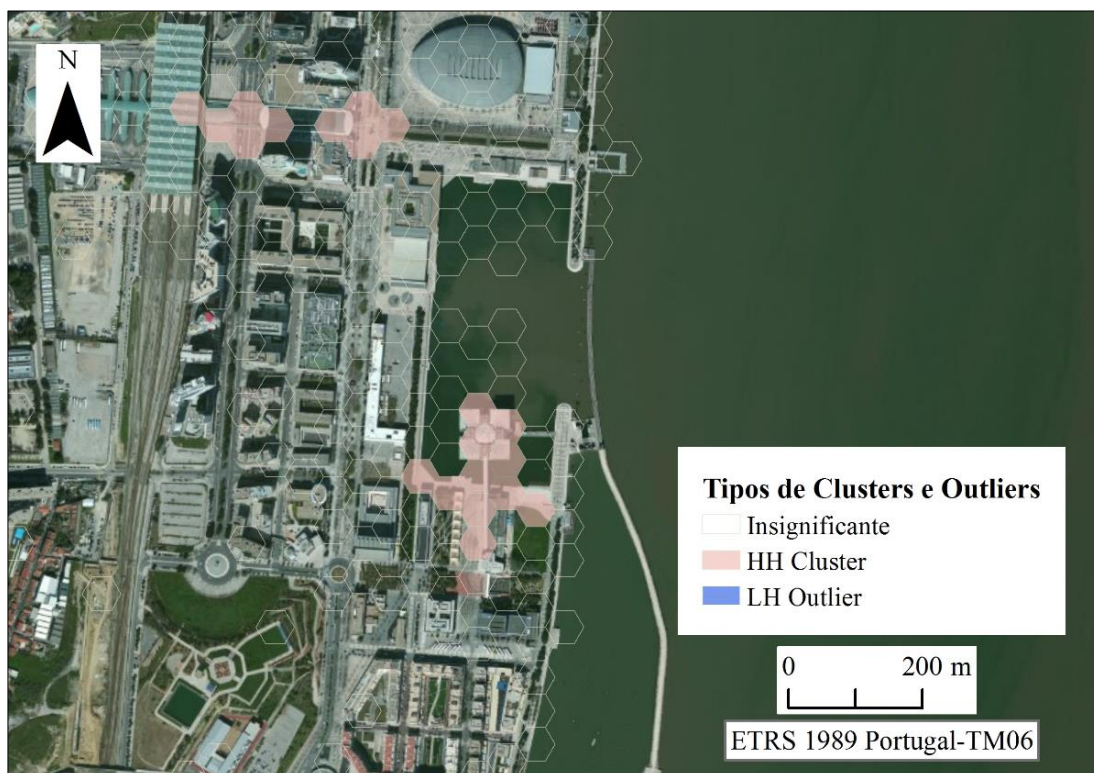


Figura 6.7: Tipos de *clusters*, zona norte.



Através da identificação dos *clusters* descobrem-se aspectos interessantes relativamente à dinâmica da distribuição espacial dos turistas de Lisboa. A dimensão da célula utilizada para a análise permitiu compreender as diferenças do padrão, nas próprias áreas que circunscrevem os *clusters* (*e.g.* na praça do Comércio, denota-se que o padrão avança pela rua Augusta apenas até a terceira intersecção, destacando a pouca relevância das ruas paralelas a ela; no Elevador de Santa Justa, a forma particular alongada do *cluster* que acompanha a rua do mesmo nome; a concentração pontual dos visitantes para registrarem fotografias desde o extremo norte do Parque Eduardo VII; ou os focos dentro do Parque das Nações).

No que respeita a significância estatística dos padrões espaciais, como se pode observar no Anexo 30, Anexo 31 e Anexo 32), os *clusters* foram identificados considerando um nível de confiança de 95% e 99% ( $z\text{-score} > 1,96$  ou  $> 2,58$  respectivamente). Contudo, deve-se referir que do total de *clusters* encontrados, a maioria pertence ao grupo cuja probabilidade de ser criado por acaso é inferior a 1%.

A significância também está suportada pelos resultados do indicador *p-value*, sempre entre 0 e 0,19, indicando que a probabilidade de resultarem de um processo aleatório é bastante reduzida.

## 6.5. Análise de pontos quentes emergentes

Antes de iniciar a discussão dos resultados da APQE, apresentam-se alguns detalhes da agregação multidimensional das fotografias dos turistas. Como foi dito no enquadramento conceptual, quando as fotografias são agregadas, a ferramenta utilizada efectua um teste de tendência geral (teste de Mann-Kendall) para verificar qual o comportamento dos dados ao longo dos anos em análise. Na Tabela 6.2 são apresentados os resultados do teste de MK (o relatório original pode ser consultado no Anexo 33).

**Tabela 6.2: Tendência dos dados agregados, no cubo de dados.**

Tendência geral dos dados	
Sentido da tendência	Decrescente
Estatística de MK	-3,7509
<i>P-value</i>	0,0002

A tendência observada nas caixas de agregação é decrescente, em concordância com a diminuição do número de observações (de 2.988 em 2007 a 862 no último ano analisado) revelada na etapa de exploração. É importante referir que a tendência avaliada no momento da agregação dos dados é estatisticamente significativa, como confirmado pelo valor de  $p$  muito baixo.

Quando a análise também é suportada pela variável tempo, denotam-se ligeiras mudanças no esquema dos locais de concentração, que decorrem, obviamente, da junção das duas componentes (espacial e temporal). Ao analisar os resultados da APQE, em comparação com os da ACO, merece destaque o surgimento de vários locais de concentração e o desaparecimento de outros.

Olhando para o mapa (Figura 6.8), identificam-se cinco locais de concentração que não tinham sido descobertos na ACO. Três deles estão fora das áreas mais turísticas da cidade (microcentralidades turísticas).

De Sul para Norte, o primeiro local (novo) situa-se sobre o Museu da Electricidade, e corresponde a um ponto que tem sido quente, nomeadamente, nos períodos mais recentes. Este aglomerado tem uma extensão de duas células com o mesmo padrão. Dentro do parque florestal de Monsanto, no antigo restaurante Panorâmico de Monsanto, desvela-se o segundo *hot spot*, que constitui uma marca pontual, sendo apenas uma célula cujo padrão temporal é de tipo consecutivo.

Mais para o interior, na freguesia de São Domingos de Benfica, sobressai o aglomerado de pontos quentes do Jardim Zoológico de Lisboa, cobrindo a quase totalidade do complexo. O padrão temporal é semelhante dos casos anteriormente reportados. Na freguesia da Estrela, localiza-se um outro aglomerado, associado à Basílica da Estrela. A mancha está confinada por duas células, com significância estatística apenas no último passo de tempo (Agosto 2013 – Julho 2014). Na freguesia de Santo António encontra-se o último destes novos focos, sobre o jardim do Torel, envolvendo a área do miradouro e constituindo um ponto quente nos intervalos de tempo finais.

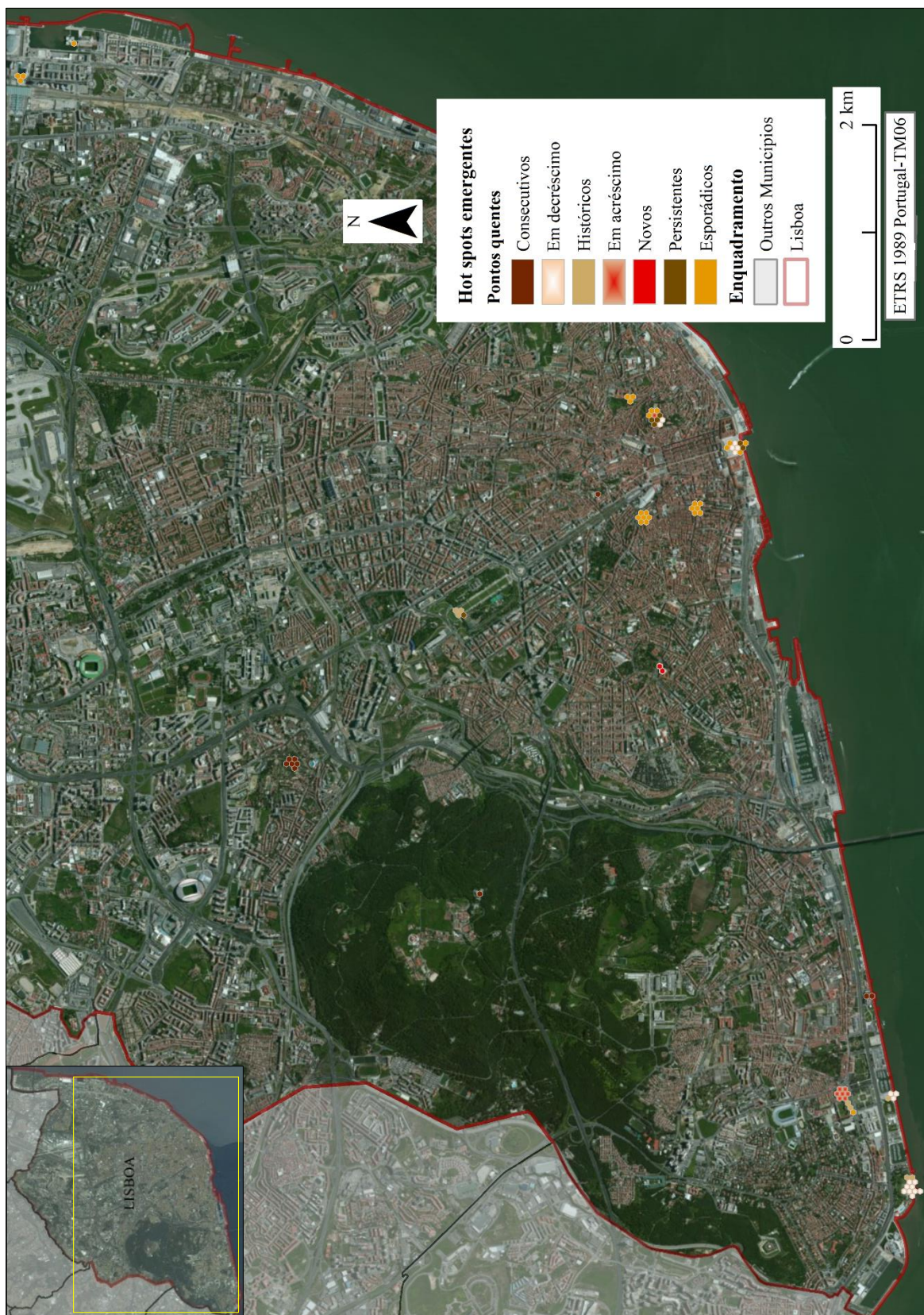
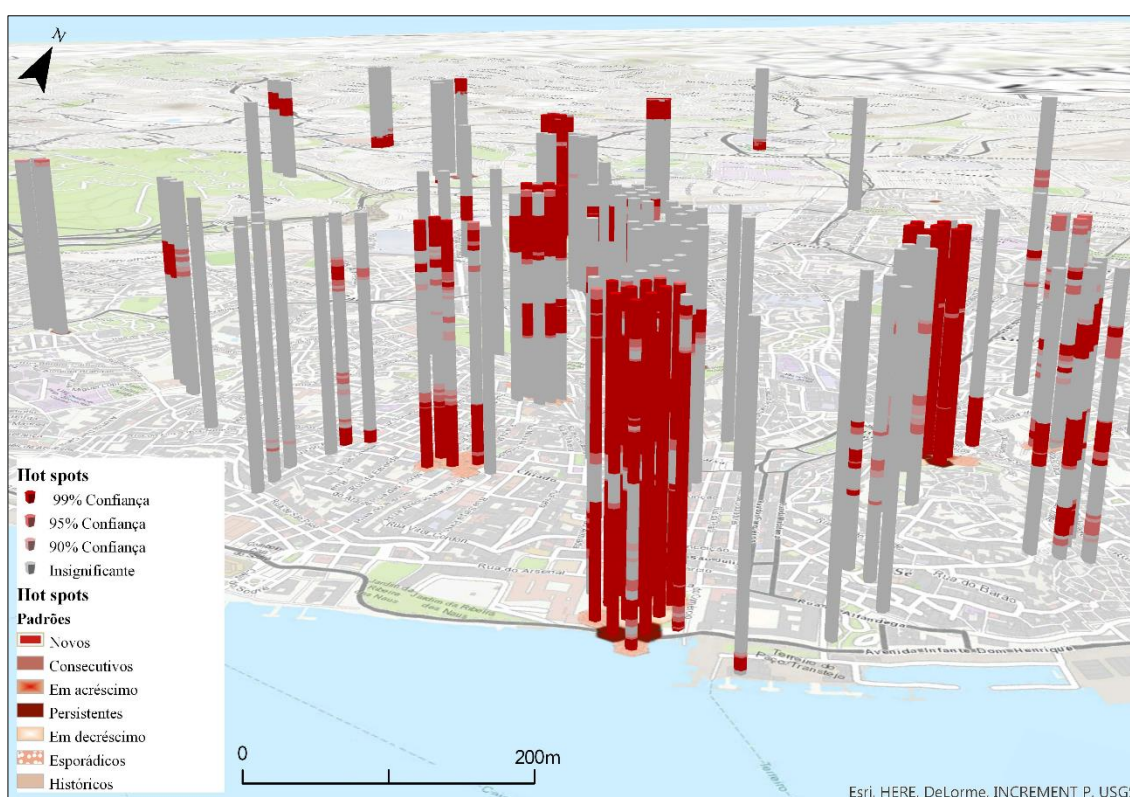


Figura 6.8: Tipos de pontos quentes emergentes.



Por outro lado, os locais de concentração que deixaram de ser significativos, no contexto espaço-temporal, encontram-se no centro da cidade. Nestes locais, cinco na freguesia de Santa Maria Maior (Miradouros das Portas do Sol e de Santa Catarina, igreja Sé de Lisboa, praças dos Restauradores e Dom Pedro IV, e Elevador de Santa Justa) e um na freguesia da Misericórdia (Elevador da Bica), a significância parece não sobressair ao longo do tempo. Por exemplo, como se pode observar na análise tridimensional (Figura 6.9), os *clusters* associados à Sé de Lisboa e ao Elevador da Bica, em muitos intervalos de tempo, chegam apenas a um nível de confiança de 90%, sendo a sua significância estatística demonstrada em meses específicos do período em análise, sem continuidade de relevo.

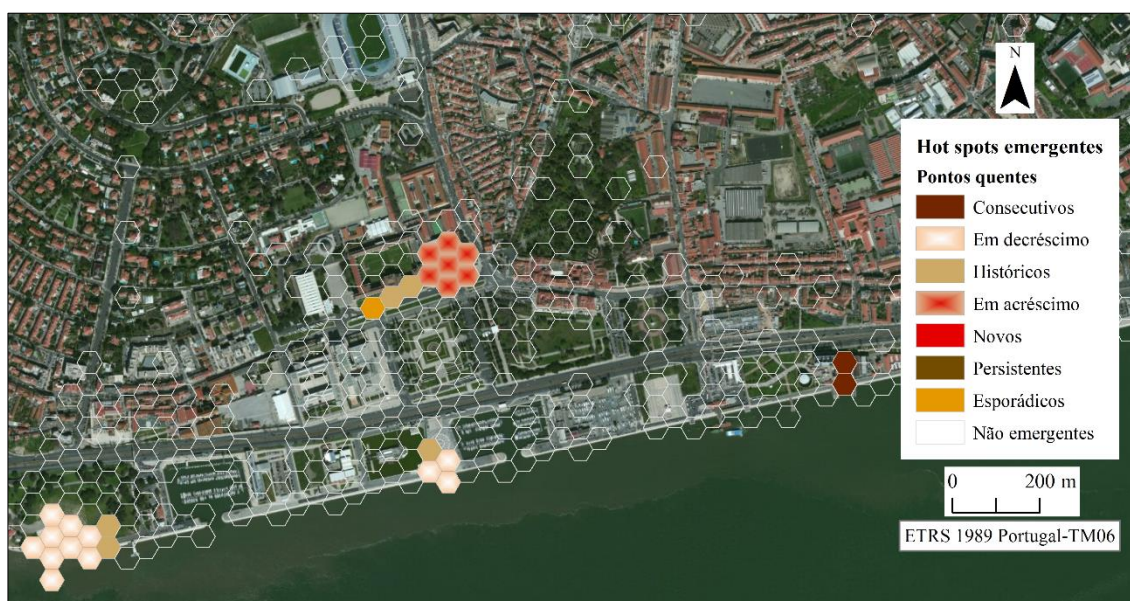


**Figura 6.9: Visualização 3D da conformação dos pontos quentes emergentes, zona centro.**

Existem também vários locais identificados na APQE que coincidem, em termos de localização, com os encontrados na ACO. Seguindo a mesma estrutura da discussão dos resultados do método ACO, os pontos quentes emergentes são apresentados, dividindo a cidade em três zonas (Sul, Centro e Norte). Para facilitar a interpretação, cada uma delas é visualizada a uma escala maior.

Na zona sul, na freguesia de Belém, estão encaixadas quatro aglomerações, uma já anteriormente descrita (museu da Electricidade) e, as restantes, associadas às atracções

do Padrão dos Descobrimentos, da Torre de Belém e do Mosteiro dos Jerónimos (Figura 6.10). Nos dois primeiros locais, é de realçar que na maioria das células que os moldam, verifica-se uma tendência temporal decrescente. Nesses locais a intensidade do padrão vem diminuindo ao longo do tempo mas sem afectar a sua significância estatística. Em ambos os aglomerados, também existem células que foram pontos quentes nos anos mais afastados, mas não no último passo de tempo (*hot spots* históricos). No caso do aglomerado associado ao Mosteiro, ele está representado por células com padrões temporais variados, *i.e.* com características muito diferentes. Enquanto na parte da entrada se verifica que a intensidade do padrão vai aumentando nos últimos anos, no resto das áreas adjacentes à fachada (em direcção ao museu da Marinha), o padrão tem representatividade nos períodos mais remotos de análise, ou apresenta um comportamento descontínuo, sendo apenas significativos em meses intervalados.

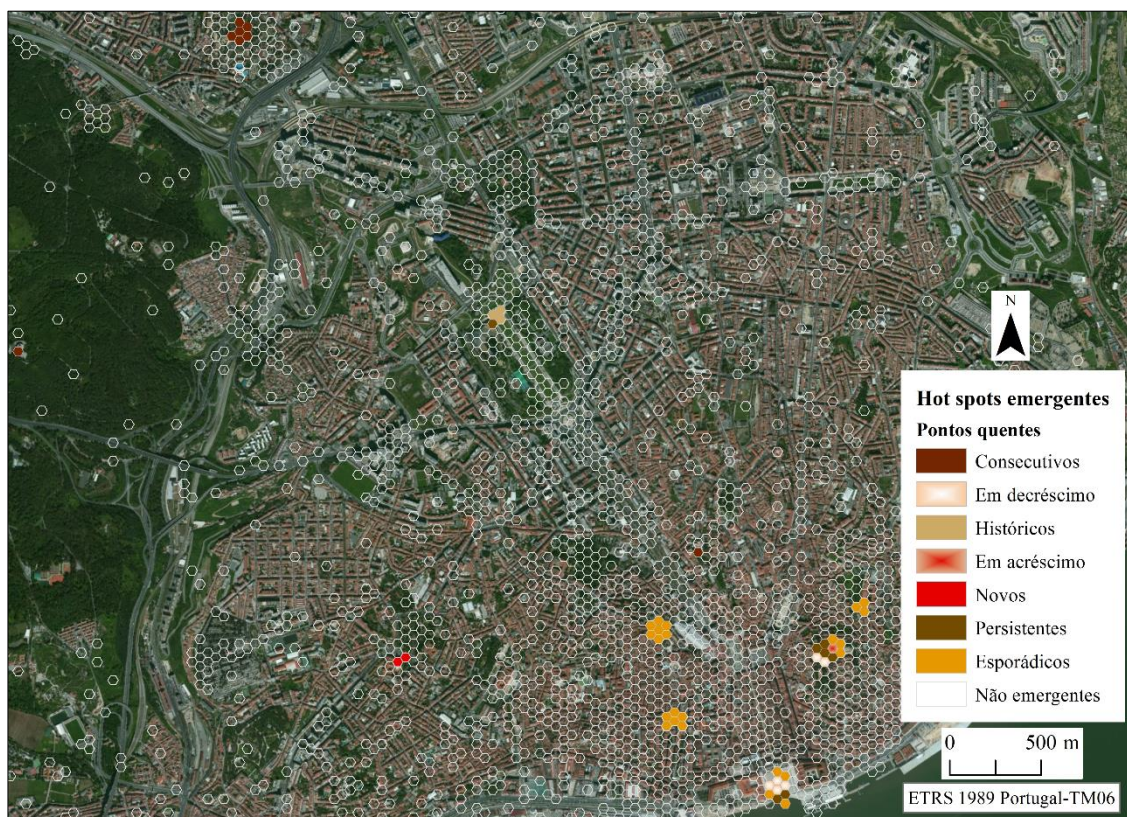


**Figura 6.10: Tipos de pontos quentes emergentes, zona sul da cidade.**

Na zona centro, identificam-se cinco locais de concentração nas freguesias da Misericórdia, Santa Maria Maior e São Vicente (Figura 6.11). Enquadradas na primeira freguesia, denotam-se duas aglomerações de pontos quentes situadas, uma sobre a praça Luís de Camões e, outra, no miradouro de São Pedro de Alcântara. Ambas se estendem um pouco mais além do perímetro dos locais de referência. A primeira chega a cobrir o Largo de Chiado e a segunda, a estação e parte do trajecto do Elevador da Glória. Nos dois casos, as suas células apresentam um padrão temporal esporádico. Enquanto na praça Luís de Camões, os pontos quentes estatisticamente significativos são detectados nos



primeiros e últimos anos, no miradouro, eles são detectados nos anos intermédios e no final do período de análise (Figura 6.9).

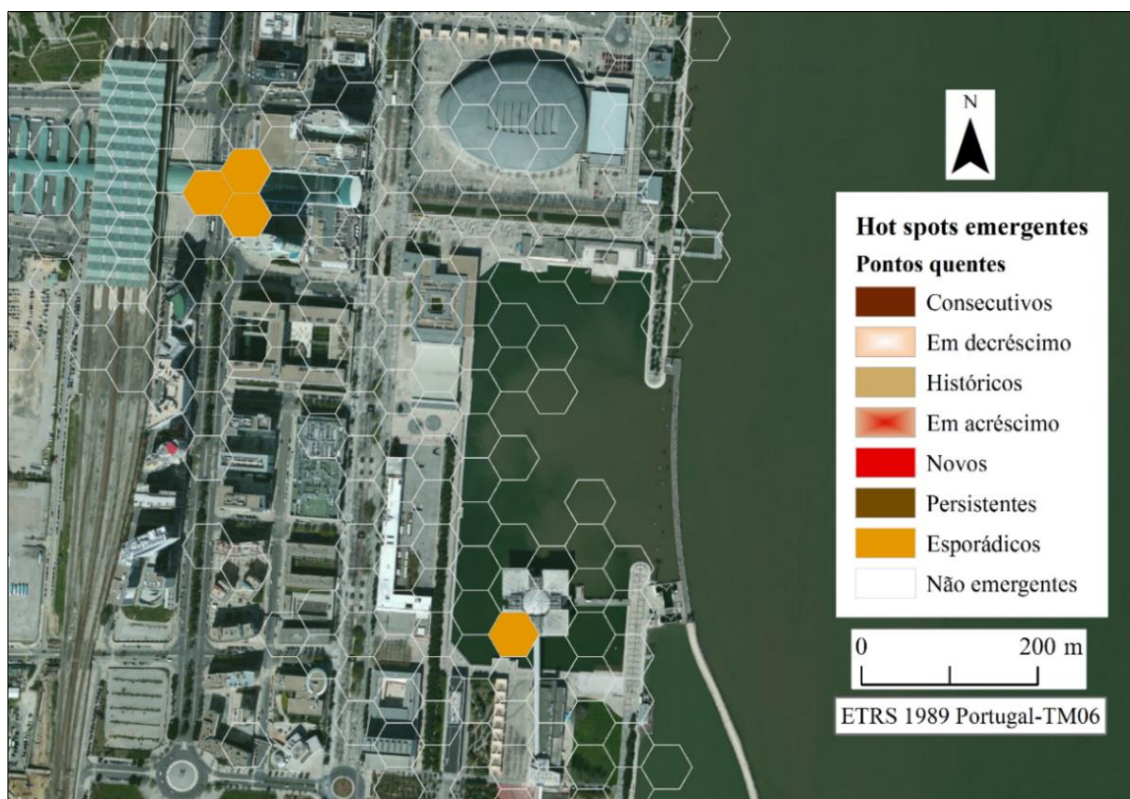


**Figura 6.11: Tipos de pontos quentes emergentes, zona centro.**

Na freguesia Santa Maria Maior, encontram-se duas aglomerações, na praça de Comércio e no Castelo de São Jorge. Na praça do Comércio, o padrão temporal do conjunto de células é diversificado, observando-se no Cais das Colunas, a existência de pontos quentes de forma continua ao longo dos anos, com subidas e descidas na intensidade do processo de *clustering*. No centro da praça, a intensidade do padrão vem a diminuir nos últimos anos e, nas fronteiras do *cluster*, encontram-se células que são estatisticamente significativas em intervalos de tempo descontínuos (Figura 6.9).

No Castelo, os padrões temporais observados nos pontos quentes também são variados, e com uma certa particularidade. Ao mesmo tempo que na fronteira do *cluster* se observa um padrão esporádico ou propenso a diminuir, nas áreas centrais, a tendência é totalmente oposta. Aqui, os pontos quentes aparecem de forma mais contínua, registando ainda, em locais específicos, um aumento da intensidade do padrão no passo de tempo final. Na freguesia das Avenidas Novas, abrangendo a área do miradouro do parque Eduardo VII, localiza-se o último dos *hot spots* desta zona, sendo sua relevância constatada nos intervalos de tempo iniciais.

Na zona norte (Figura 6.12), o número de locais relevantes é menor que nas outras zonas, concentrando-se no complexo do Parque das Nações. São dois os locais existentes, um associado a uma atracção turística (o Oceanário de Lisboa) e, o outro, a um serviço de apoio (a estação de comboio de Oriente).



**Figura 6.12: Pontos quentes emergentes, zona norte.**

Quando comparados com os resultados da ACO, nota-se uma redução da dimensão dos *clusters* de ambos. No primeiro caso, apenas persiste a célula próxima da entrada do edifício e, no segundo caso, só permanecem as células situadas entre a estação de comboios e o centro comercial Vasco da Gama. Os pontos quentes estatisticamente significativos em ambos os locais vão surgindo de forma esporádica no decorrer da série temporal em análise.

Confirma-se a importância destes locais, que sempre se mostraram como focos da actividade nas diferentes tentativas de análise.

Num registo um pouco diferente pode-se ainda fazer uma análise espaço-temporal incidindo na sequência de locais visitados e não na frequência. Efectivamente, recorrendo à sequência de carregamento de fotos que cada utilizador realiza é possível identificar o percurso realizado. Claro que, como já foi referido, o utilizador pode não carregar as fotografias no exacto momento em que está a realizar o

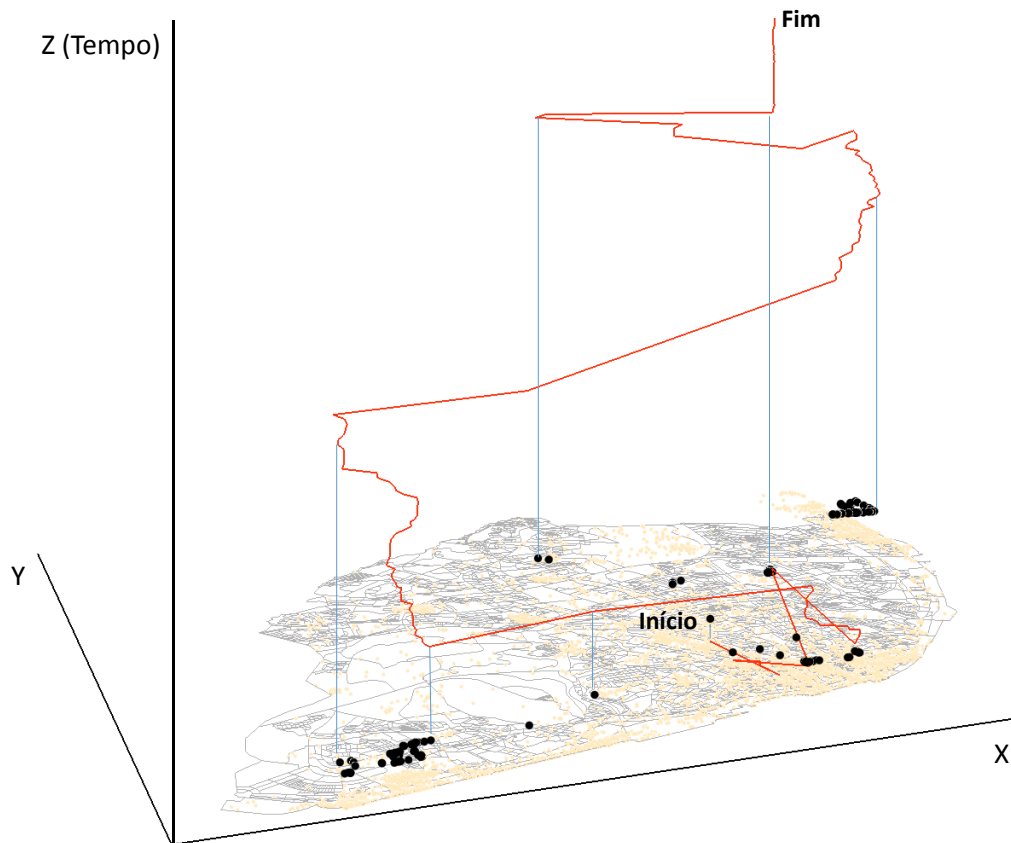
percurso, mas estando a operar-se no capítulo da '*Big Data*', manuseando milhares de registos, algum erro que possa daí advir é estatisticamente desprezível.

Deve-se também ter em atenção que a base de dados do Panoramio apenas indica a sequência de carregamento de fotos, utilizando somente a data, ao contrário de outras redes como o Flickr, que recorrem à data-hora. Assim sendo, apenas é possível identificar o local para onde o turista se deslocou mas não quanto tempo demorou no trajecto. Igualmente, tratando-se de carregamentos de fotos, o trajecto gerado traduz a distância euclidiana entre os locais sucessivamente visitados e não o percurso efectivo tomado pelo utilizador da rede.

Na Figura 6.13, encontra-se representado o percurso diário de um turista típico de Lisboa. O ponto de partida situa-se na microcentralidade identificada como centro histórico, o local onde há maior oferta hoteleira (apesar do Parque das Nações também ter vindo a adquirir notoriedade neste campo). Depois de visitar vários locais no centro o utilizador dirigiu-se para Belém e depois ao Parque das Nações, regressando de seguida ao ponto de partida. Deste modo foram visitadas três microcentralidades, que podem muito bem ser quatro, se for assumido – o que não é de todo improvável – que o percurso entre a baixa e Belém, por exemplo, tenha sido efectuado junto à zona Ribeirinha.

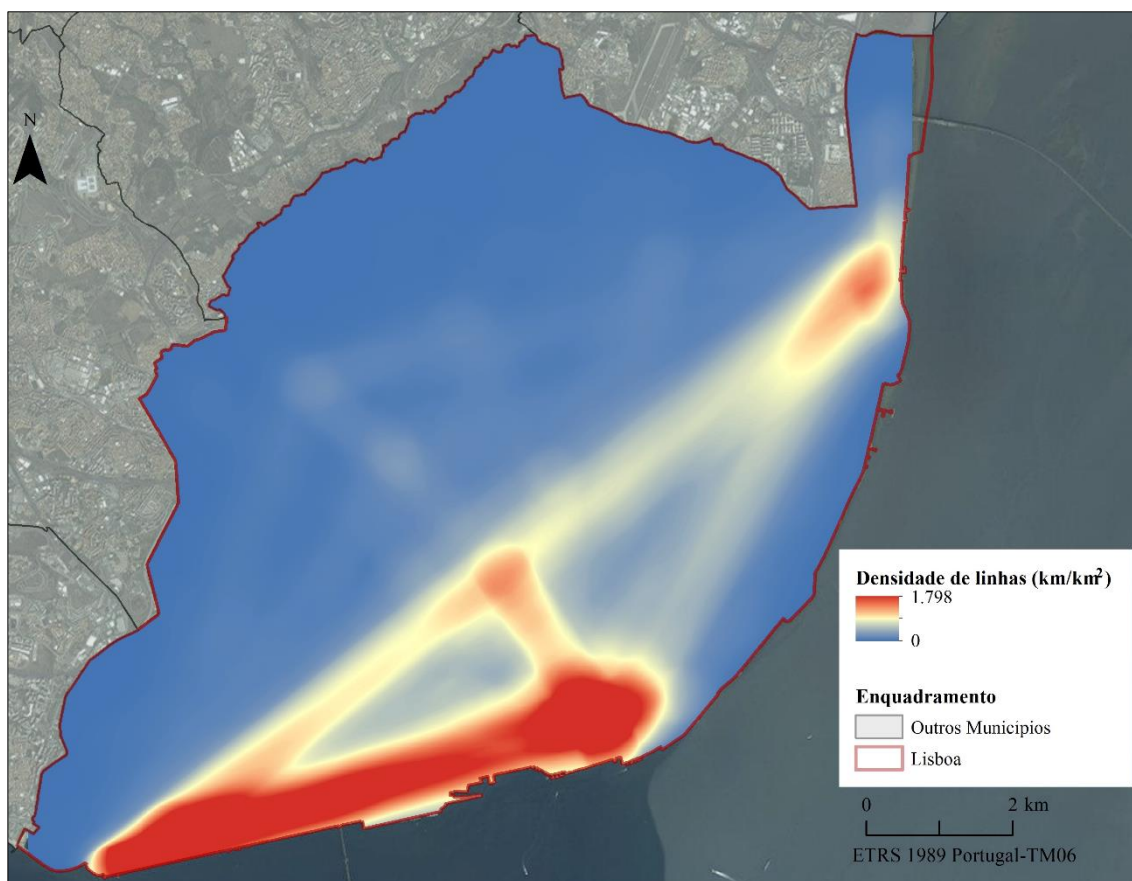
Pelos pontos correspondentes às fotografias tiradas, também se pode inferir que os locais de atracção no centro histórico se encontram bastante mais dispersos do que em Belém e no Parque das Nações, onde o grau de aglomeração é notório. Paralelamente, o número de fotografias tiradas fora destas microcentralidades é residual, o que é indicador que, mesmo no caso do eixo ribeirinho, este funciona como um corredor de acesso e não como um real ponto de atracção.





**Figura 6.13: 24 horas na vida de um turista de Lisboa.**

A análise da Figura 6.14 corrobora a análise até agora feita. Esta imagem corresponde a uma densidade de linhas que, tal como o nome indica, corresponde à densidade de elementos lineares na vizinhança de cada célula em análise, expressa em unidades de comprimento por unidade de área ( $\text{km}/\text{km}^2$ ). Os dados utilizados foram todos os percursos realizados pelos turistas durante o período em análise. Observando-se os fluxos – assim chamados por não corresponderem a percursos reais – dos turistas, facilmente se depreende que os grandes focos de atracção são o centro histórico e Belém. A ligação faz-se, preferencialmente, de forma directa ao longo do arco ribeirinho, existindo no entanto, um comportamento alternativo, que corresponde à utilização do eixo Av. Liberdade – Parque Eduardo VII, como vértice de um triângulo de ligação. O Parque das Nações assume-se como um ponto importante mas com ligações relativamente fracas ao triângulo anteriormente referido.



**Figura 6.14: Densidade de fluxos de turistas em Lisboa.**

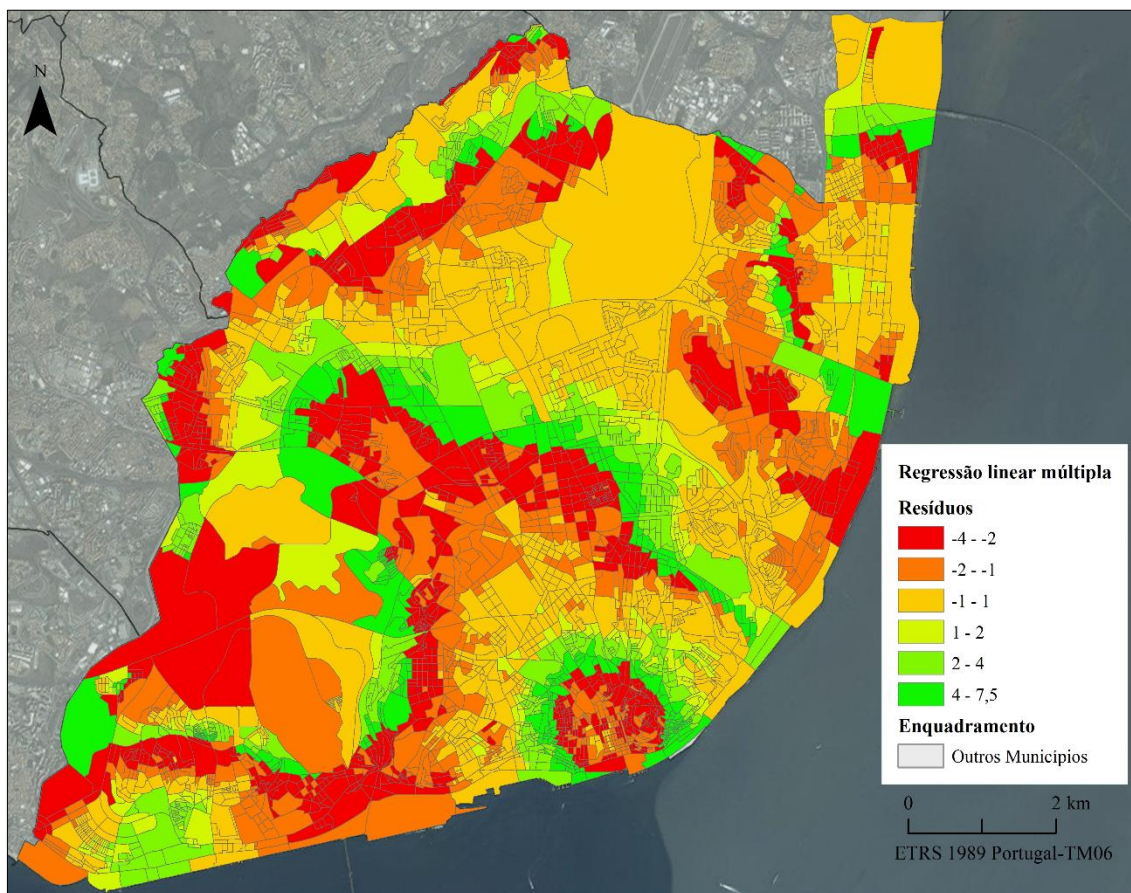
## 6.6. Análise do potencial territorial

Da análise de regressão linear múltipla, retirou-se que os factores mais importantes na explicação do número de fotografias tiradas eram (Tabela 5.6) a proximidade a: Monumentos de Interesse Público (0,29); Docas de Recreio e Marinas (0,21); Arquitectura Premiada (0,15); Geomonumentos (0,12); Hotelaria (0,10); e Teatros (0,10). No entanto, mais do que a análise dos factores e da predição do modelo, o que pode ser mais interessante em termos de planeamento turístico é uma análise dos resíduos do modelo.

Efectivamente, na Regressão Linear Múltipla as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. A Análise dos Resíduos é precisamente um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos seus resíduos. Como visto anteriormente, o resíduo é dado pela diferença entre a variável resposta observada e a variável resposta estimada. A ideia base da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem reflectir as propriedades impostas pelo termo de erro do modelo. Mais importante, os

resíduos positivos indicam resultados acima do predito pelo modelo e resíduos negativos indicam o inverso.

Na prática, analisando a Figura 6.15, o que a informação acerca dos resíduos indica é que, face às condições observadas, *i.e.*, factores explicativos, há zonas em que o número de fotos é superior ao expectável (resíduos positivos) e outras onde há menos fotos do que seria previsível. São precisamente estas áreas, que podem necessitar de intervenções no âmbito do planeamento para que possam explanar todo o seu potencial.



**Figura 6.15: Resíduos do modelo de Regressão Linear Múltipla.**

Na Figura 6.15, para além de Belém sobressair com valores acima do que o seu potencial justificaria, destaca-se de forma inesperada a zona da Baixa, que apesar de em todas as análises anteriores aparecer como um importante *hot spot*, ainda se encontra abaixo do que potencialmente poderá atingir. A zona do Parque das Nações segue a mesma tendência mas não com um diferencial tão grande. Contudo, são claramente duas zonas, duas microcentralidades, identificadas pelo gabinete de turismo da CML, que podem ser alvo de acções, tendo em vista a captação de mais visitantes.



## 7. Conclusões

Nesta dissertação, tirou-se proveito das fotos *geotagged* partilhadas pelos utilizadores da rede Panoramio, levantadas durante a sua visita a Lisboa entre 2007 e 2014, para analisar o comportamento espaciotemporal da procura turística na cidade. Este tipo de informação (*crowdsourced data*), proveniente de redes sociais, aporta à compreensão da fruição/consumo do espaço dentro dos destinos turísticos urbanos. Particularmente, os dados extraídos do Panoramio podem fornecer informação minuciosa, de grande valor, para a identificação dos locais de concentração dos turistas, em áreas densas e complexas.

Cada vez que um turista tira uma foto, está a fazer uma selecção entre todos os elementos que possivelmente poderiam ser fotografados. A foto digital transformou os hábitos fotográficos dos turistas, denotando-se um claro aumento do número de fotos tiradas. Além disso, as imagens adoptaram uma função social, pois são frequentemente compartilhadas em redes sociais, *i.e.*, têm uma componente viral muito maior. Ao mesmo tempo, as fotografias partilhadas em redes sociais permitem o acesso a um catálogo muito extenso de imagens que anteriormente seriam muito difíceis de obter.

As fotografias tiradas por turistas tornam possível identificar e analisar as áreas de atracção turística visual numa cidade. A maioria dos estudos baseados em informação geolocalizada proveniente de redes sociais têm-se limitado a criar produtos de visualização, *e.g.* mapas de pontos e animações. A presente dissertação representa um avanço face a esse estado da arte, visto que as fotografias do Panoramio foram utilizadas para criar mapas de *clusters* espaço-temporais, estatística multivariada e espacial.

Os resultados mostram que os padrões de distribuição observados nas fotografias obtidas por turistas e por residentes apresentaram tanto semelhanças como diferenças. As fotografias dos turistas mostraram uma maior concentração espacial do que as obtidas pelos residentes. A comparação entre os *clusters* espaciais das fotos dos turistas e os mapas dos principais pontos de atracção (incluindo as microcentralidades) indicam que os resultados são consistentes, uma vez que os clusters espaciais coincidiram com áreas da cidade que são conhecidas por atrair mais turistas.

O conjunto de dados (de 17.604 observações utilizadas para a análise) constitui uma amostra robusta que replica o espaço vivido pelos turistas. Os resultados da avaliação estatística da tendência geral (global) de *clustering* evidenciou a sua aptidão para a análise

espacial, deixando de lado qualquer possibilidade que sua génese seja produto da aleatoriedade.

As fotografias dos turistas – não obstante o reconhecimento de que os utilizadores do Panoramio terão, por certo, um perfil que não espelha com absoluto rigor e exaustão as características dos turistas que visitam a cidade - são um indicador muito relevante da imagem percebida de um destino. Os dados obtidos da rede, em termos gerais, aparentam algumas semelhanças com outras estatísticas oficiais utilizadas para “medir” o turismo urbano como por exemplo, o Inquérito às actividades dos turistas e a informação levantado pelo Observatório de Turismo de Lisboa, nomeadamente os números apresentados pelas taxas de ocupação hoteleira.

No caso de Lisboa, o mapeamento e consequente análise das fotografias levaram a constatar que, efectivamente, os turistas tendem a se concentrar nas áreas com maior potencial turístico da cidade (definidas como microcentralidades turísticas pela administração local). As áreas mais salientes estão bem demarcadas, sendo frequentemente percorridas, em conjunto, pelos turistas: Belém, centro histórico (estendendo-se até ao Marquês de Pombal) e Parque das Nações. São os elementos monumentais os que se mostram como focos de maior atractividade dos turistas, (*e.g.* castelo, praças, igrejas). Os espaços verdes tais como parques e jardins, em muitos casos, ligados aos elementos supramencionados, são também locais alvos dos turistas. Denota-se a importância, por um lado, da faixa ribeirinha, do centro para o sul e, por outro lado, dos locais elevados (miradouros) situados nas colinas da cidade. Em menor grau mas constituindo parte da oferta turística, a infra-estrutura de apoio, sobretudo de transporte, apresenta alguma relevância quando esta é simplesmente um serviço (*e.g.* Estação de comboios do Oriente) ou quando vê-se complementada como um produto turístico (*e.g.* Elevador da Glória).

Consequentemente, estudos futuros, usando fotografias georreferenciadas de redes sociais como um *proxy* para a distribuição dos turistas nas cidades podem concentrar-se em fotografias tiradas por turistas usando procedimentos de segmentação, não sendo necessário comparar padrões espaciais de turistas e residentes. Efectivamente, uma vantagem das *geotagged* fotos é que podem ser utilizadas como um *proxy* para a distribuição espacial dos turistas, permitindo medir, cartografar e fazer comparações dentro das cidades.

A análise de uma amostra representativa de fotografias tiradas em Lisboa permitiu determinar as características mais relevantes das fotografias tiradas pelos turistas de visita à cidade. Na verdade, os fotógrafos/turistas não compartilham todas as suas imagens, preferem fazer uma selecção premeditada, optando por compartilhar algumas e descartar outras. O resultado é um filtro duplo da sua percepção da cidade, em primeiro lugar, através da selecção do local da obtenção e, em seguida, pela escolha do conjunto de imagens carregadas. A análise de imagens compartilhadas em redes sociais contribui para estudos de turismo, pois permite traçar as rotas dos turistas com base nos indícios fotográficos que partilham e, posteriormente, comparar os padrões de distribuição e mobilidade com o conteúdo e perspectivas das imagens obtidas.

As áreas de interesse podem ser definidas analisando a concentração das fotografias dos turistas através de operações realizadas em SIG. Algumas delas podem nem ser consideradas atracções turísticas pelas autoridades competentes. Desta forma, podem saber quais as áreas que são mais atractivas para os visitantes e quais são as características dessas áreas (espaços verdes, monumentos, lugares com arquitectura moderna, etc.), aspectos que são fundamentais para comercializar os destinos.

Torna-se claro que estas análises podem ser muito úteis para estudar a imagem percebida pelos turistas e também podem ser aplicadas ao marketing e gestão dos pontos de destino. É também interessante notar que as fotos carregadas no Panoramio não representam todas as fotografias capturadas pelos turistas, mas apenas aquelas que realmente foram seleccionadas para a partilha. Embora haja alguma evidência que imagens do Panoramio coincidem com as fotos que os turistas tiram nas suas viagens, seria interessante estudar a relação entre as fotos tiradas por turistas em viagem e as colocadas na internet.

Outro potencial interessante na análise das fotografias representa a sucessão temporal das imagens dos utilizadores. Os locais podem atrair um número distinto de pessoas em diferentes períodos de tempo. A inclusão da componente temporal na análise, para além de revelar a evolução dos locais de concentração já identificados, resulta no descobrimento de novos locais que não parecem relevantes considerando exclusivamente o contexto espacial. O padrão temporal observado nos locais de concentração associados às atracções mais relevantes da cidade (*e.g.* Torre de Belém, Mosteiro dos Jerónimos, Praça de Comercio, Castelo de São Jorge) não muda significativamente ao longo dos

anos, constatando-se, na maioria das células que conformam os aglomerados, a continuidade do processo espacial (*clustering*) mas diferindo apenas na sua intensidade.

Uma vez que as fotos dos turistas estão indexadas conforme os seus respectivos *timestamps*, podem contribuir para o conhecimento dos fluxos de turistas dentro do ambiente urbano. Na sequência da análise dos fluxos, verifica-se que os focos de atracção mais relevantes, o centro histórico e Belém, estão conectados, preferencialmente, pelo arco ribeirinho. Como alternativa, a ligação entre ambos os focos é também dada através do eixo Av. Liberdade – Parque Eduardo VII. O Parque das Nações, no norte, aparece como um ponto importante mas com ligações relativamente fracas.

A distribuição espacial dos turistas não é homogénea, na tentativa de esclarecer quais os factores que promovem a variação do processo espacial, confirma-se que são os monumentos de interesse público (em conjunto com mais quatro factores: Docas de recreio e Marina; Arquitectura Premiada; Geomonumentos; Hotelaria e Teatros), os elementos que melhor conseguem explicar a aglomeração geográfica das fotografias. Da experiência com as variáveis seleccionadas, apesar de se explicar apenas 68% da variação total do fenómeno, deve-se considerar o aporte que a leitura dos resíduos da regressão linear múltipla traz para a análise do potencial territorial. Segundo o modelo de regressão, existem áreas dentro das microcentralidades (i) Baixa-Chiado e (ii) Alfama-Castelo-Mouraria cujo potencial turístico está subestimado. Estas áreas podem necessitar de intervenções no âmbito do planeamento, seja do espaço público, actividades, ou prestação de serviços, de modo a acertar o volume de turistas que a sua “capacidade” concede.

A compreensão dos padrões de movimento intra-destino dos turistas revelou-se extremamente interessante. O modelo foi desenvolvido dedutivamente com base na sequência de carregamento das fotografias por parte de cada utilizador. As implicações da pesquisa nesta área podem ser significativas, especialmente para os destinos que são altamente dependentes do turismo e onde os turistas compõem uma grande parte do movimento pedestre e de transportes públicos. Existem três áreas em que o conhecimento de movimentos intra-destino pode ter um impacto importante, nomeadamente o planeamento de transportes, o desenvolvimento de produtos e da imagem, e a gestão do impacto do turismo.

No que se refere ao planeamento de transporte, o conhecimento dos itinerários diários dos turistas e os fatores que influenciam esses itinerários, pode ajudar os fornecedores de transporte a atender as necessidades de forma mais eficiente, bem como



a coordenar melhor as viagens com os fluxos de transporte local. A informação acerca do movimento também pode ser usada para identificar os afunilamentos e entraves desnecessários no fluxo entre os locais de alojamento e as atracções ou outros destinos.

Relativamente ao desenvolvimento de produtos. Sabendo quais as preferências dos turistas, podem-se identificar quais os caminhos e destinos que podem ser usados para melhor definir as atracções existentes, planejar alternativas e comercializá-las de forma mais eficaz. O conhecimento das rotas dos turistas pode ser usado para definir os limites de microcentralidades e nós (*hot spots*). Esta informação pode ser usada para desenvolver novas atracções e produtos ao longo das rotas comuns e nas microcentralidades e nós de destino.

Os impactos do turismo tendem a ocorrer quando os turistas frequentam um lugar mais do que ele pode sustentar. O conhecimento, no tempo e no espaço, das rotas mais frequentadas e dos destinos mais visitados- bem como as variações de intensidade ao longo do ciclo circadiano de cada atracção turística -, pode ser usado para identificar os períodos e os locais que excederem a sua capacidade e têm o potencial de causar efeitos sócio ambientais negativos, ou impactos culturais. Alternativamente, também podem ser identificados padrões de subutilização, e podem ser desenvolvidos planos de gestão para mitigar os impactos negativos deslocando a utilização de rotas e locais, dos períodos mais críticos para horários, caminhos e lugares alternativos.

Os padrões específicos da cidade têm implicações importantes para a indústria do turismo e para as autoridades. Os profissionais de hotelaria, donos de lojas e restaurantes, podem identificar as áreas de maior concentração turística para planejar os seus negócios em conformidade. As empresas responsáveis por serviços e produtos de turismo também podem saber com maior acuidade onde instalar os seus recursos. Ao fazer isto, os planeadores saberão quais os pontos de maior concentração, de forma a identificar os espaços com maior grau de saturação. Isto irá permitir o controlo da capacidade de carga que tenham em consideração a conservação dos locais, a segurança e a convivência salutar entre turistas e locais.

Estes resultados também são importantes para os turistas, pois esta informação é deveras útil para planejar a viagem e tomar outro tipo de decisões associadas. Especificamente, saber de antemão as dimensões da cidade e o grau de dispersão-concentração dos pontos turísticos, ajuda a determinar o total de tempo alocado à visita, a seleccionar a acomodação de acordo com a proximidade a locais de interesse, a contratar

serviços de turismo (guias, autocarros [*sightseeing buses*], etc.) e até mesmo comprar bilhetes de transporte específicos para turistas, de acordo com a duração da estada, com o tipo/preferência quanto ao meio de deslocação, ou mesmo na criação de uma oferta integrada entre transporte público e entradas em atracções turísticas seleccionadas.

Para finalizar, deve-se apontar aquela que é considerada por muitos a principal limitação deste tipo de estudos e, consequentemente, desta dissertação. Em primeiro lugar, nem todos os turistas que visitam a cidade tiram fotos e/ou as publicam *on-line*. Portanto, este não é um estudo de todos os turistas, mas apenas dos turistas fotógrafos que compartilham as suas fotos na rede Panoramio.

Assim, a informação é, efectivamente, parcial, na medida em que nem todos os turistas fazem uso destas redes, e aqueles que o fazem, fazem-no com diferentes graus de intensidade. Além disso, a informação refere-se exclusivamente a locais visitados pela maioria dos turistas e por eles considerados esteticamente atraentes, não reflectindo totalmente a atracção de outros locais menos "fotogénicos". Isto afecta particularmente os locais visualmente indistinguíveis como locais de negócios, estudo e compras, que podem estar sub-representados. Por exemplo, para o turismo de compras, algumas áreas comerciais (mercados de rua) aparecem nessas fontes de dados mas outras como centros comerciais (onde é proibido tirar fotografias) não. Uma das formas possíveis de tentar corrigir este enviesamento em trabalhos futuros poderá ser através da comparação desta informação com dados provenientes do uso de meios de pagamento automático (*e.g.* cartão de crédito).

No entanto, é inegável que o turista/fotógrafo faz uma selecção, dentro de todas as fotografias tiradas, daquelas que pretende partilhar com as outras pessoas. Tendo em conta que os meios clássicos de divulgação das imagens turísticas (guias de viagem, brochuras, etc.) estão em competição com as novas plataformas de redes sociais, a investigação futura deve concentrar-se numa análise aprofundada dos processos internos que fazem os indivíduos seleccionar certas fotos para carregar na Internet; bem como identificar as suas características sociodemográficas e as diferenças entre os utilizadores e não utilizadores desta ferramenta de comunicação.

Considerando os prós e os contras, parece inegável que as fotografias disponibilizadas nas redes sociais podem ser um meio para obter assinaláveis progressos na análise e investigação geográfica aplicada ao turismo.

## Bibliografia

- Aldstadt J (2010) Spatial Clustering, in *Handbook of Applied Spatial Analysis Software Tools Methods and Applications*, 279-298. Fischer M, Getis A (Eds.). Springer – Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03647-7
- Andrienko G, Andrienko N, Bak P, Kisilevich S, Keim D (2009) Analysis of community-contributed space and time referenced data (example of Flickr and Panoramio photos). *Visual Analytics Science and Technology*, 2009, 213-214. Institute of Electrical and Electronics Engineers (IEEE Computer Society). doi: 10.1109/VAST.2009.5333472
- Andrienko N, Andrienko G (2006) Exploratory analysis of spatial and temporal data. A systematic approach, 17-20. Springer - Verlag Berlin Heidelberg. ISBN-13 978-3-540-25994-7
- Anselin L (1994) Exploratory spatial data analysis and geographic information systems, in *New tools for spatial analysis*, 45–54. Painho M (Ed.). Luxembourg, Eurostat. ISBN 92-826-8400-8
- Anselin L, Syabri I, Kho Y (2010) GeoDa: An Introduction to Spatial Data Analysis, in *Handbook of Applied Spatial Analysis Software Tools Methods and Applications*, 73-87. Fischer M, Getis A (Eds.). Springer – Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03647-7
- Anselin Luc (1995) Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2):93-115. Ohio State University Press.
- Antoniou V, Morley J, Haklay M (2010) Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1):99-110. <https://www.researchgate.net/publication/235044899>
- Arctur D, Zelier M (2004) Designing geodatabases. Case studies in GIS data modelling ESRI Press. ISBN 1-58948-021-X
- Ashworth G (2012) Do we understand urban tourism? *Journal of Tourism and Hospitality* 1(4):1-2. ISSN: 2167-0269, an open access journal. <http://www.omicsgroup.org/journals/2167-0269/2167-0269-1-e117.pdf>
- Ashworth G, Tunbridge J (2000) The Tourist-Historic City: Retrospect and prospect of managing the heritage city. Elsevier science Ltd. ISBN 0-08-043675-7

- Bivand R, Pebesma E, Gómez-Rubio V (2008) Applied spatial data: analysis with R. Gentleman R, Hornik K, Parmigiani G (Eds.). Springer Science+Business Media, LLC. doi: 10.1007/978-0-387-78171-6
- Brito Henriques E (1996) Lisboa turística. Entre o imaginário e a cidade. Lisboa: Edições Colibri. ISBN: 9789728288297
- Bruns A (2007) Produsage: Towards a broader framework for User-led content creation. Published in: *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*, 99-105. doi: 10.1145/1254960.1254975
- Brunsdon C (2008) Exploratory spatial data analysis (ESDA), in *Encyclopedia of geographic information science*, 136-140. Kemp K (Ed.). SAGE Publications. ISBN 978-1-4129-1313-3
- Buhalis D, Amaranggana A (2014) Smart tourism destinations. *Proceedings of the international conference: Information and communication technologies in tourism 2014*, 553-564. Xiang Z, Tussyadiah L (Eds). ISBN 978-3-319-03973-2
- Buhalis D, Law R (2008) Progress in information technology and tourism management: 20 years on and 10 years after the internet - The state of eTourism research. *Tourism Management*, 29:609-623. doi: 10.1016/j.tourman.2008.01.005
- Charlton M (2008) Quantative methods and geographic information systems, in *The handbook of Geographic Information Science*, 379-394. Wilson J, Fotheringham A (Eds.). Blackwell Publishing, Ltd. ISBN 978-1-4051-0795-2
- Clark P, Evans F (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445-453. Consultado em Março, 2016, em <http://www.jstor.org/stable/1931034>
- Cox D, Stuart A (1955) Some quick tests for trend in location and dispersion. *Biometrika*, 42:80-95. [https://www.jstor.org/stable/2333424?seq=1#page\\_scan\\_tab\\_contents](https://www.jstor.org/stable/2333424?seq=1#page_scan_tab_contents)
- Darcy S, Small J (2008) Theorizing precincts: disciplinary perspectives, in *City Spaces – Tourist Places: Urban Tourism Precincts*, 63-91. Hayllar B, Griffin T, Edwards D (Eds.) Elsevier Ltd. ISBN: 978-0-7506-8195-7
- Diaz L, Granell C, Huerta J, Gould M (2012) Web 2.0 Broker: A standards-based service for spatio-temporal search of crowd-sourced information. *Applied Geography*, 35:448-459. <http://dx.doi.org/10.1016/j.apgeog.2012.09.008>

- Diggle P, Gatrell A, Bailey T, Rowlingson B (1996) Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers, New Series*, 21(1):256-274. Blackwell Publishing on behalf of The Royal Geographical Society (with the Institute of British Geographers). ISSN 0020-2754
- Direção-Geral do Território (2012) Carta Administrativa Oficial de Portugal (CAOP) 2012. Versão 2012.1. Consultado em Fevereiro, 2016. Disponível em: [http://www.dgterritorio.pt/cartografia\\_e\\_geodesia/cartografia/carta\\_administrativa\\_oficial\\_de\\_portugal\\_\\_caop\\_/caop\\_\\_download\\_/carta\\_administrativa\\_oficial\\_de\\_portugal\\_\\_\\_versao\\_2012\\_1/](http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_administrativa_oficial_de_portugal__caop_/caop__download_/carta_administrativa_oficial_de_portugal___versao_2012_1/)
- Dixon P (2002) Ripley's *K* Function, in *Encyclopedia of Environmetrics*, 3:1796–1803. El-Shaarawi A e Piegorisch W (Eds.). John Wiley & Sons, Ltd. ISBN 0471 899976
- Donaire J, Camprubí R, Galí N (2014) Tourist clusters from Flickr travel photography. *Tourism Management Perspectives*, 11:26-33. doi: 10.1016/j.tmp.2014.02.003
- Dumbrovská V, Fialová D (2014) Tourist intensity in capital cities in central Europe: Comparative analysis of tourism in Prague, Vienna and Budapest. *Czech Journal of Tourism*, 3:5-26. doi: 10.2478/cjot-2014-0001
- Edwards D, Griffin T, Hayllar B (2008) Urban tourism research: developing an agenda. *Annals of Tourism Research*, 35(4):1032–1052. doi: 10.1016/j.annals.2008.09.002
- Entidade Regional de Turismo em Lisboa (2015) Plano Estratégico para o Turismo na Região de Lisboa [PETRL] 2015 – 2019. Descarregado em Fevereiro, 2016. Disponível em: <http://www.ertlisboa.pt/pt/institucional/plano-estrategico-para-o-turismo-na-regiao-de-lisboa-2015-2019/>
- Fischer M (1999) Spatial analysis: retrospect and prospect, in *Geographical Information Systems. Principles, Techniques, Management and Applications*, 1:283-292. Longley P, Goodchild M, Maguire D, Rhind D (Eds.). John Wiley and Sons. ISBN 0471–33132–5
- Fotheringham A, Brunson C, Charlton M (2000). Quantitative geography: Perspectives on spatial data analysis. SAGE Publications, Ltd. ISBN 978-0-7619-5947-2

- García-Palomares J, Gutiérrez J, Mínguez C (2015) Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, 63:408-417. doi: 10.1016/j.apgeog.2015.06.002
- Getis A (1999) Spatial statistics, in *Geographical Information Systems. Principles, Techniques, Management and Applications*, 1:283-292. Longley P, Goodchild M, Maguire D, Rhind D (Eds.). John Wiley and Sons. ISBN 0471-33132-5
- Getis A (2010) Spatial Autocorrelation, in *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, 255-278. Fischer M e Getis A (Eds.). Springer – Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03647-7\_14
- Getis A, Aldstadt J (2004) Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2):90-104. The Ohio State University.
- Getis A, Franklin J (1987) Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68:473-477. Ecological Society of America, Washington DC.
- Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):190-206. Ohio State University Press.
- Gilbert R (1987) Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold Company, Inc. ISBN 0-442-23050-8
- Girardin F, Fiore F, Ratti C, Blat J (2008) Leveraging explicitly disclosed location information to understand tourist dynamics: A Case Study. *Journal of Location Based Services*, 2(1):41-56. <http://dx.doi.org/10.1080/17489720802261138>
- Goodchild M (2007) ‘Citizens as sensors: the world of volunteered geography’, *GeoJournal*, 69:211-221. doi: 10.1007/s10708-007-9111-y
- Goodchild M (2008) Assertion and authority: The science of User-generated geographic content’, in *Proceedings of the Colloquium for Andrew U. Frank’s 60th Birthday*. <http://www.geog.ucsb.edu/ygood/papers/454.pdf>
- Goodchild M (2008) Data Analysis, Spatial, in *Encyclopedia of GIS*, 200-204. Shekar S, Xiong H (Eds.) Springer Science + Business Media, LLC. ISBN 978-0-387-35973-1

- Goodchild M, Haining R (2004) GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83:363-385. doi: 10.1007/s10110-003-0190-y
- Goodchild M, Li L (2012) Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110-120. doi: 10.1016/j.spasta.2012.03.002
- Gospodini A (2001) Urban Design, Urban Space Morphology, Urban Tourism: An Emerging Paradigm Concerning Their Relationship. *European Planning Studies* 9(7):925–934. doi: 10.1080/09654310120079841
- Graham P (2005) Web 2.0. Consultado em Janeiro 2016, url: <http://www.paulgraham.com/web20.html>
- Hanning R (2004) Spatial data analysis: Theory and practice. Cambridge University Press. ISBN 0-511-04085-7
- Hawelka B, Sitko I, Beinatz E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41:260-271. doi: 10.1080/15230406.2014.890072
- Hayllar B, Griffin T (2005) The precinct experience: a phenomenological approach. *Tourism Management*, 26:517–528. doi: 10.1016/j.tourman.2004.03.011
- Hayllar B, Griffin T, Edwards D (2008) Urban Tourism Precincts: Engaging with the field, in *City Spaces – Tourist Places: Urban Tourism Precincts*, 3-18. Hayllar B, Griffin T, Edwards D (Eds.) Elsevier Ltd. ISBN: 978-0-7506-8195-7
- Heilmann C, Francis N (2007) Web development solutions, Ajax, APIs, Libraries, and Hosted Services made easy. Springer-Verlag New York, Inc. ISBN: 1-59059-806-7
- Herrera A, San Martín H, Hernández J (2015) Perceived influence on behavior of user-generated content on social network sites: An empirical application in the hotel sector. *Revista Española de Investigación de Marketing ESIC*, 19:12-23. <http://dx.doi.org/10.1016/j.reimke.2014.09.001>
- Hetland M (2008) Beginning python from novice to professional, Second Edition. Springer-Verlag New York, Inc. ISBN: 1-59059-982-9

- Hirsch R, Slack J (1984) A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, 20(6):727-732. doi: 10.1029/WR020i006p00727
- Hirsch R, Slack R., Smith R (1982) Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18(1):107-121. doi: 10.1029/WR018i001p00107
- Hubert G, Costanzo C (1981) Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 13(3):224-233. Ohio State University Press.
- Illian J, Pettinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns, 425-437. John Wiley & Sons Ltd. ISBN 978-0-470-0149-2
- Instituto Geográfico Português (2012) Metadados da CAOP 2012.0. Instituto Geográfico Português, Ministério da agricultura, do mar, do ambiente e do ordenamento do território. Descarregado em Fevereiro, 2016. Disponível em: [http://ftp.igeo.pt/produtos/cadastro/caop/download/METADADOS\\_CAOP20121.pdf](http://ftp.igeo.pt/produtos/cadastro/caop/download/METADADOS_CAOP20121.pdf)
- Instituto Nacional de Estatística, I.P. (2012) Censos 2011 Resultados Definitivos - Região Lisboa. INE - Portugal. ISBN 978-989-25-0185-7. Descarregado em Fevereiro, 2016. Disponível em: [http://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=ine\\_censos\\_publicacoes](http://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=ine_censos_publicacoes)
- Instituto Nacional de Estatística, I.P. (2015) Estatísticas do Turismo 2014. INE – Portugal. ISBN 978-989-25-0308-0. Descarregado em Fevereiro, 2016. Disponível em: [https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine\\_publicacoes&PUBLICACOESpub\\_boui=139601&PUBLICACOESmodo=2&xlang=pt](https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=139601&PUBLICACOESmodo=2&xlang=pt)
- Jacquez G (2008) Spatial cluster analysis, in *The handbook of Geographic Information Science*, 396-416. Wilson J, Fotheringham A (Eds.). Blackwell Publishing, Ltd. ISBN 978-1-4051-0795-2



- Judd D (2003) Visitors and the spatial ecology of the City, in *Cities and Visitors: regulating people, markets, and city space*, 13-38. Hoffman L, Fainstein S, Judd D (Eds.). Blackwell Publishing Ltd.
- Kádár B (2014) Measuring tourist activities in cities using geotagged photography, *Tourism Geographies*, 16(1):88-104. doi: 10.1080/14616688.2013.868029
- Kádár B, Gede M (2013) Where do tourists go? Visualizing and analysing the spatial distribution of geotagged photography. *Cartographica*, 48(2):78-88. University of Toronto Press. doi: 10.3138/carto.48.2.1839
- Kaplan A, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, (53):59-68. doi: 10.1016/j.bushor.2009.09.003
- Kendall M (1975) Rank correlation methods, 4th edition, Charles Griffin, London. ISBN 9780852641996
- Khaliq M, Ouarda T, Gachon P, Sushma L, St-Hilaire A (2009) Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers. *Journal of Hydrology*, 368:117–130. <http://dx.doi.org/10.1016/j.jhydrol.2009.01.035>
- Kisilevich S, Keim D, Andrienko N, Andrienko G (2013) Towards acquisition of semantics of places and events by multi-perspective analysis of geotagged photo collections. *Geospatial Visualisation*, 211-233. doi: 10.1007/978-3-642-12289-7\_10
- Kisilevich S, Krstajic M, Keim D, Andrienko N, Andrienko G (2010) Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. *IV 2010 Information and Visualization*, 289-296. Institute of Electrical and Electronics Engineers (IEEE Computer Society). doi: 10.1109/IV.2010.94
- Koerbitz W, Önder I, Hubmann-Haidvogel A (2013) Identifying tourist dispersion in Austria by digital footprints. *Proceedings of the international conference: Information and communication technologies in tourism 2013*, 495-506. Cantoni L, Xiang Z (Eds). ISBN 978-3-642-36309-2

- Lau G, McKercher B (2006) Understanding tourist movement patterns in a destination: A GIS approach. *Tourism and Hospitality Research*, 7(1):39-49. Sage Publications, Ltd.
- Lee G, Yamada I, Rogerson P (2010) GeoSurveillance: GIS-based exploratory spatial analysis tools for monitoring spatial patterns and clusters, in *Handbook of Applied Spatial Analysis Software Tools Methods and Applications*, 135-150. Fischer M, Getis A (Eds.). Springer – Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03647-7
- Lee J, Wong D (2001) Statistical analysis with ArcView GIS. John Wiley & Sons, Inc. ISBN 0-471-34874-0
- Leung D, Law R, Van Hoof H, Buhalis D (2013) Social media in tourism and hospitality: A literature review, *Journal of Travel & Tourism Marketing*, 30:3-22. doi: 10.1080/10548408.2013.750919
- Lew A, McKercher B (2006) Modelling tourist movements: A local destination analysis. *Annals of Tourism Research*, 33(2):403–423. doi: 10.1016/j.annals.2005.12.002
- Lloyd D (2010) Spatial Data Analysis, an introduction for GIS users. Oxford University Press Inc., New York. ISBN 978-0-19-955432-4
- Lo I, Mckercher B, Lo A, Cheung C, Law R (2011) Tourism and online photography. *Tourism management*, 32:725-731. doi: 10.1016/j.tourman.2010.06.001
- Longley P, Goodchild M, Maguire D, Rhind D (2005) Geographical Information Systems and Science - 2nd Edition. John Wiley & Sons Ltd. ISBN 0-470-87000-1
- Lu X, Wang C, Yang J, Pang Y, Zhang L (2010) Photo2trip: Generating travel routes from geo-tagged photos for trip planning. *Proceedings of the international conference on multimedia*, 143-152. doi: 10.1145/1873951.1873972
- Mackay K, Vogt C (2012) Information technology in everyday and vacation contexts. *Annals of Tourism Research*, 39(3):1380-1401. <http://dx.doi.org/10.1016/j.annals.2012.02.001>
- Mann H (1945) Non-parametric tests against trend. *Econometrica*, 13:163-171. [http://www.jstor.org/stable/1907187?seq=1#page\\_scan\\_tab\\_contents](http://www.jstor.org/stable/1907187?seq=1#page_scan_tab_contents)

- Maruch S e Maruch A (2006) Python for dummies. John Wiley & Sons. ISBN: 9780471778646
- Mas J, Mateu C (2010) Introduction to web application development, p:5-17.
- Miller H (2004) Tobler's First Law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284-289. Blackwell Publishing, Malden, MA.
- Mitchell A (1999) The ESRI guide to GIS analysis. Volume 1: Geographic patterns & relationships. ESRI Press. ISBN 9781879102064
- Mitchell A (2005). The ESRI guide to GIS analysis, Volume 2: Spatial measurements and statistics. ESRI Press. ISBN: 9781589481169
- O'Sullivan D, Unwin D (2010) Geographic information analysis. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN: 978-0-470-28857-3
- Observatório de Turismo de Lisboa, OTL (2014) Inquérito às actividades dos turistas e informação Região de Lisboa 2014. OTL. Descarregado em Fevereiro, 2016. Disponível em: <http://www.visitlisboa.com/Observatorio.aspx>
- Observatório de Turismo de Lisboa, OTL (2014) Inquérito Motivacional Cidade de Lisboa 2014. OTL. Descarregado em Fevereiro, 2016. Disponível em: <http://www.visitlisboa.com/Observatorio.aspx>
- Openshaw S (1983) The modifiable areal unit problem, concepts and techniques in modern geography, 3-41. ISBN-10: 0860941345
- Page S (1995) Urban tourism. Routledge. ISBN 0-415-11218-4
- Pearce D (2001) An integrative framework for urban tourism research. *Annals of Tourism Research*, 28(4):926-946. PII:S0160-7383(00)00082-7
- Peuquet D (1994) It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441-461. doi: 10.1111/j.1467-8306.1994.tb01869
- Plano Regional de Ordenamento do Território da área Metropolitana de Lisboa, PROTAML (2009) Diagnóstico Sectorial Sistema Urbano, Habitação e Equipamentos Constitucional. Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo (CCDR LVT). Descarregado em Fevereiro,

2016. Disponível em: <http://www.ccdr-lvt.pt/pt/plano-regional-de-ordenamento-do-territorio-da-area-metropolitana-de-lisboa/54.htm>

Rodrigues A (2010) Income distribution dynamics in The European Union: Exploring (In)stability measures within geographical clusters. *50th Anniversary - European Congress of the Regional Science Association International*. <http://www.fcsh.unl.pt/~egeo/?q=node/1826>

Rodrigues A, Tenedório J (2014) Sensitivity analysis of spatial autocorrelation using distinct geometrical settings: Guidelines for the urban econometrician. *Computational Science and Its Applications – ICCSA 2014*, 345-356.

Rodrigues A, Tenedório J (2016) Sensitivity analysis of spatial autocorrelation using distinct geometrical settings: Guidelines for the quantitative geographer. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, 7(1):65-77. doi:10.4018/IJAEIS.2016010105

Scott L, Getis A (2008) Spatial statistics, in *Encyclopedia of geographic information science*, 436-440. Kemp K (Ed.). SAGE Publications. ISBN 978-1-4129-1313-3

Scott L, Janikas M (2010) Spatial statistics in ArcGIS, in *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, 27-40. Fischer M, Getis A (Eds.). Springer – Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03647-7

Sharpley R, Roberts L (2005) Part III: Managing tourism in its environment - Managing urban tourism. *The Management of Tourism*, 161-174. ISBN 0761940219

Sheller M, Urry J (2004) Tourism mobilities: Place to play, places in play. Routledge. ISBN 0-203-34033-7

Shoval N e Isaacson M (2010) Tourist mobility and advanced tracking technologies. Routledge. ISBN 0-203-86937-0

Straumann R, Çöltekin A, Andrienko G (2014) Towards (Re)constructing narratives from georeferenced photographs through visual analytics. *The Cartographic Journal*, 51(2):152-165. doi: 10.1179/1743277414Y.0000000079

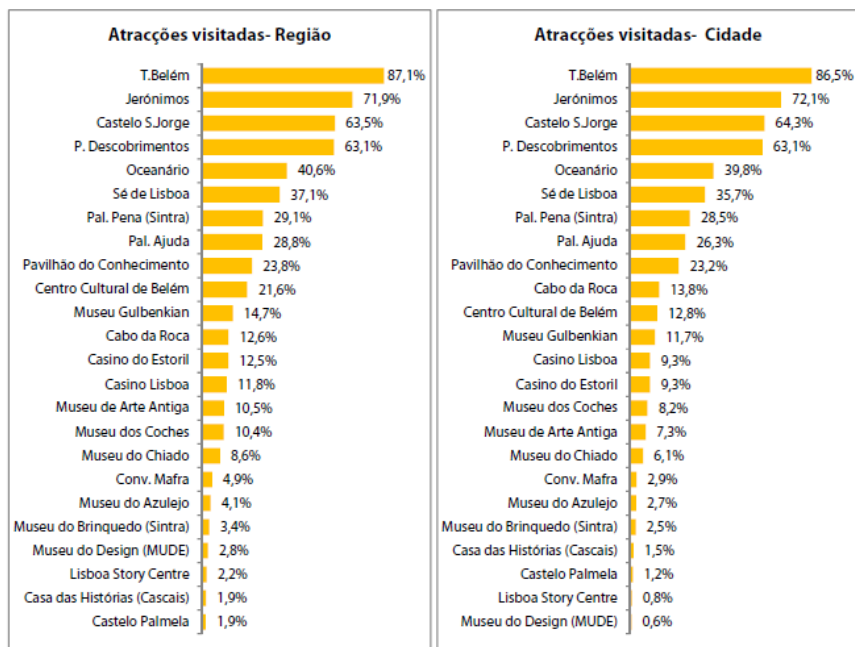
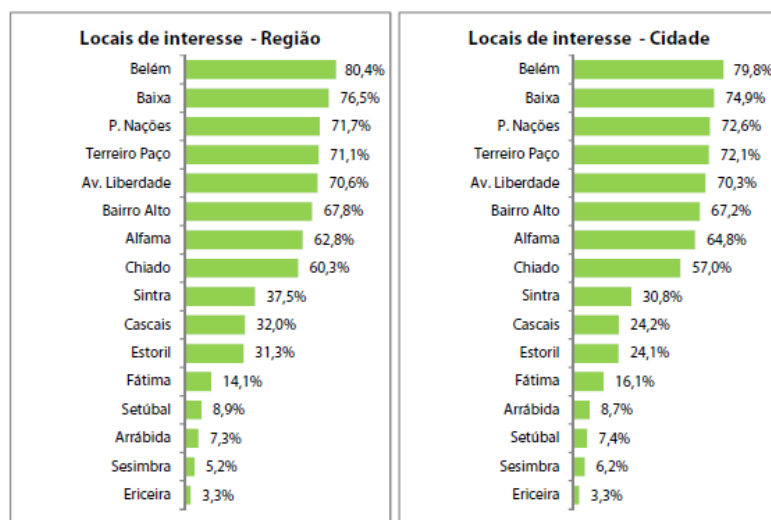
Sui D, Goodchild M (2011) The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11):1737-1748. <http://dx.doi.org/10.1080/13658816.2011.604636>

- Sui D, Goodchild M, Elwood S (2013) Volunteered Geographic Information, the exaflood, and the growing digital divide, in *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, 1-12. Sui D, Elwood S, Goodchild M (Eds). ISBN 978-94-007-4587-2
- Tammet T, Luberg A, Järv P (2013) Sightsmap: Crowd-sourced popularity of the world places. *Information and Communication Technologies in Tourism 2013*, 314-325. Cantoni L, Xiang Z (Eds.). ISBN 978-3-642-36308-5
- Tobler W (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography. Proceedings of the International Geographical Union. Commission on Quantitative Methods*, 46:234-240.
- Tussyadiah L (2012) An assessment of contagion on social networking sites. *Proceedings of the international conference: Information and communication technologies in tourism 2012*, 25-35. Fuchs M, Ricci F, Cantoni L (Eds). ISBN 978-3-7091-1141-3
- Unwin D (2008) Spatial Analysis, in *Encyclopedia of geographic information science*, 392-396. Kemp K (Ed.). SAGE Publications. ISBN 978-1-4129-1313-3
- Urry J, Larsen J (2011) The tourists gaze 3.0., p. 155-188. London: SAGE Publications Ltd. ISBN 978-1-84920-377-7
- Van der Knaap W (1999) Research report: GIS-oriented analysis of tourist time-space patterns to support sustainable tourism development, *Tourism Geographies*, 1(1):56-69. doi: 10.1080/14616689908721294
- Wakefield J, Lyons H (2010) Spatial aggregation and the Ecological Fallacy, in *Handbook of spatial statistics*, 541-558. Gelfand A, Diggle P, Fuentes M, Guttorp P (Eds.). Taylor and Francis Group, LLC. ISBN 978-1-4200-7287-7
- Waller L, Gotway C (2004) Applied spatial statistics for public health data. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN 0-471-38771-1
- Yue S, Pilon P, Cavadias G (2002) Power of the Mann-Kendall and Spearman's rho test for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259:254–271. [http://dx.doi.org/10.1016/S0022-1694\(01\)00594-7](http://dx.doi.org/10.1016/S0022-1694(01)00594-7)

Zielstra D, Hochmair H (2013) Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science*, 58(2):251-273.  
doi: 10.1080/14498596.2013.801331

## Anexos

### Anexo 1. Locais de interesse e atracções visitadas. Fonte: Inquérito às actividades dos turistas e informação 2014, Região de Lisboa. Observatório de Turismo de Lisboa.





## **Anexo 2. Script original para retirar informação da rede de partiha de fotos Panoramio.**

```
import requests
import json
import os

search_name = "BellingcatRaqqah"
southwest_corner = "35.92865398664048,38.96073818206787"
northeast_corner = "35.97895687940326,39.06510829925537"

# we use the coordinates
lat_min,long_min = southwest_corner.split(",")
lat_max,long_max = northeast_corner.split(",")

# Function responsible for sending requests to Panoramio API

def send_panoramio_request(start,end):

    api_url = "http://www.panoramio.com/map/get_panoramas.php?set=full&"
    api_url += "from=%d&to=%d&" % (start,end)
    api_url += "minx=%s&miny=%s&maxx=%s&maxy=%s&" %
(long_min,lat_min,long_max,lat_max)
    api_url += "size=medium&mapfilter=false"

    response = requests.get(api_url)

    if response.status_code == 200:

        # convert to a dictionary
        search_results = json.loads(response.content)

        return search_results

    # there was a problem
    return None

# Use the Panoramio API to get all pictures

def get_all_panoramio_pictures():

    photo_list = []
    search_start = 0

    # send the initial request
    search_results = send_panoramio_request(search_start,search_start+50)

    # if there was an error return nothing
    if search_results is None:
        print "[*] Error retrieving photos."
        return

    # add the current results to our list
    photo_list.extend(search_results['photos'])

    print "[*] Retrieved %d photos..." % (len(photo_list))
```

```

# while there are more photos to retrieve
while search_results['has_more'] is True:

    # we increase the search count by 50
    search_start += 50

    search_results = send_panoramio_request(search_start,search_start+50)

    if search_results is not None:
        photo_list.extend(search_results['photos'])

        print "[*] Retrieved %d photos..." % (len(photo_list))

# return all of our photos
return photo_list

# Write out the list of photos

def write_photo_list(photo_list):

    if not os.path.exists("%s" % search_name):
        os.mkdir("%s" % search_name)

    fd = open("%s/%s.html" % (search_name,search_name),"wb")
    fd.write("<html><head></head><body>")

    # walk through the list of photos and add them to our log file
    for photo in photo_list:
        fd.write("<a target='_blank' href='https://maps.google.com/?q=%f,%f'><img src='%s'
border='0'></a><br/>\r\n" %
                (photo['latitude'],photo['longitude'],photo['photo_file_url']))

    # close the html file
    fd.write("</body></html>")
    fd.close()

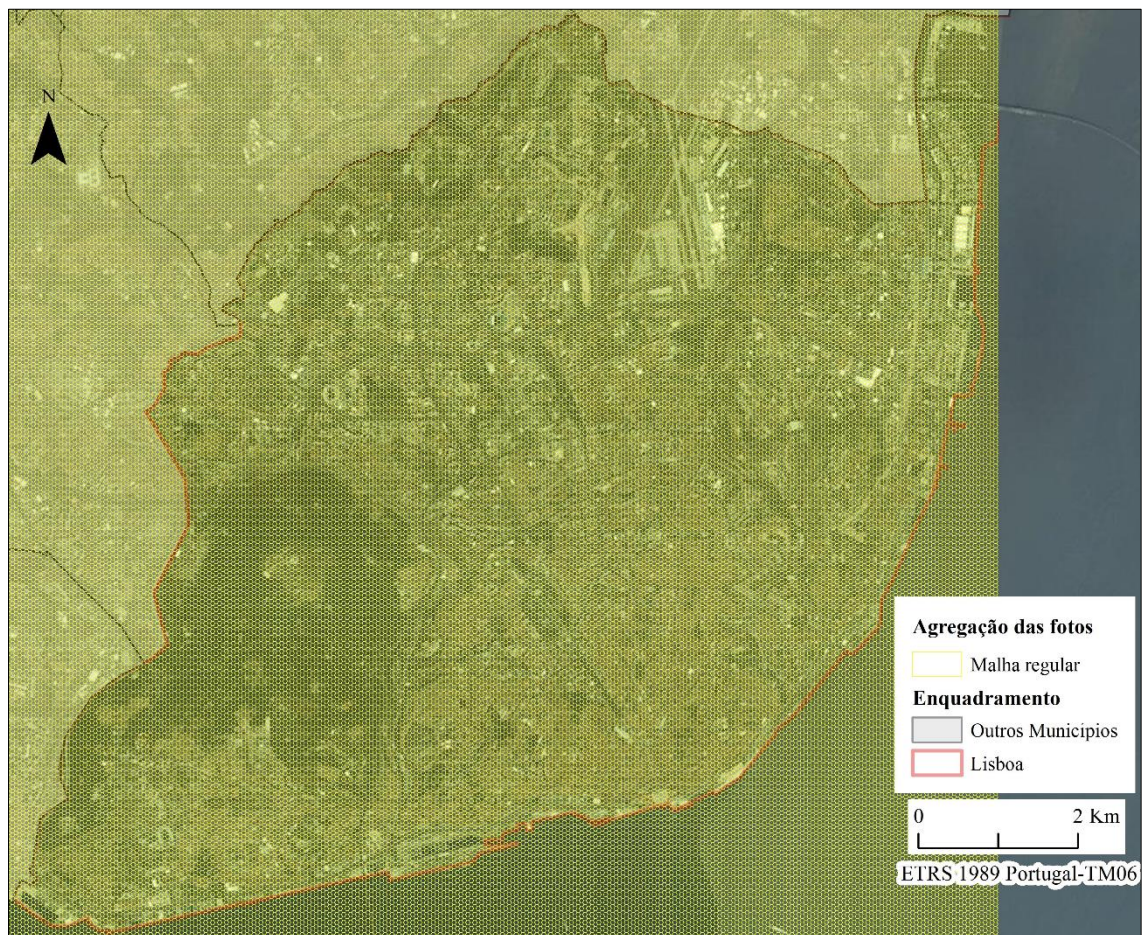
    return

# grab the full list of photos from panoramio
photo_list = get_all_panoramio_pictures()

write_photo_list(photo_list)

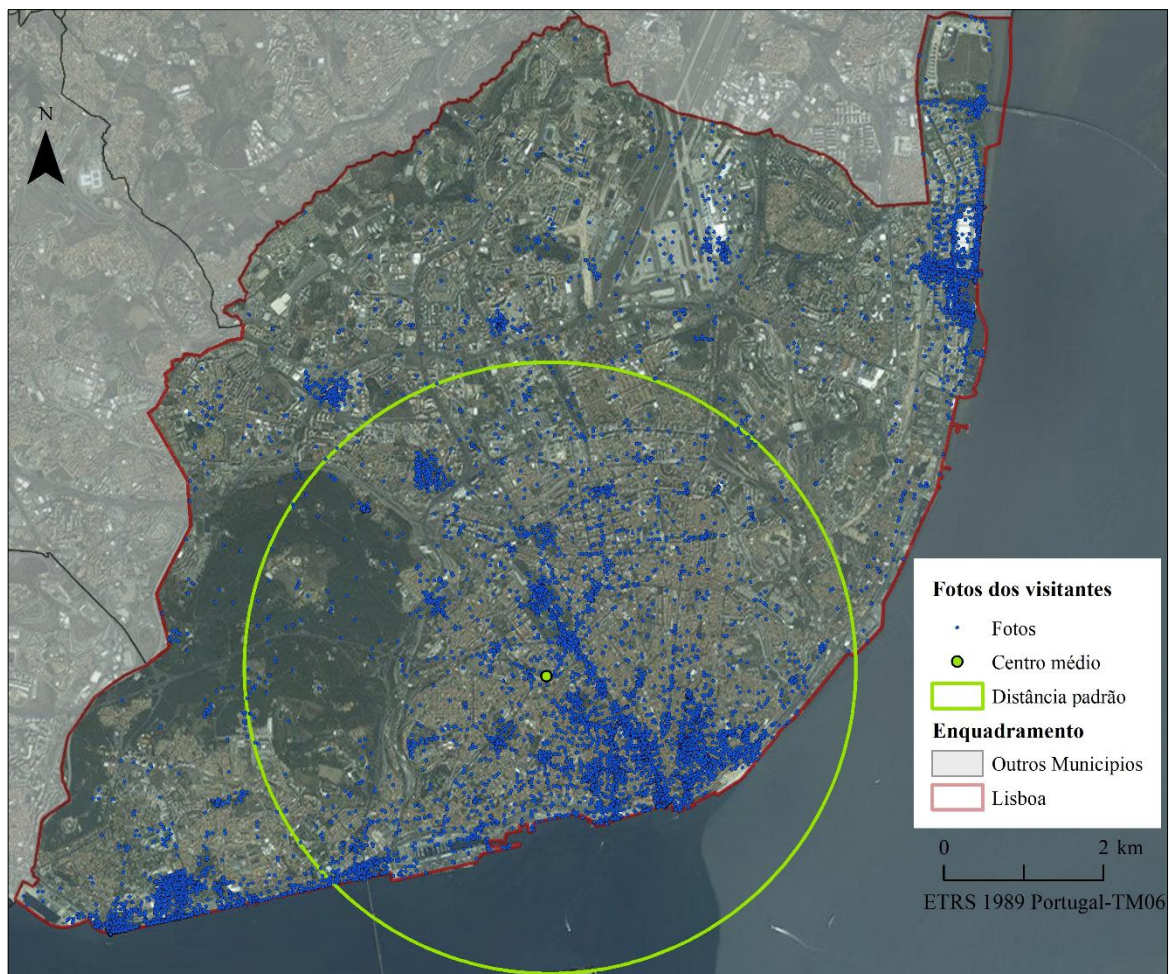
```

**Anexo 3. Malha regular de hexágonos, criada a partir da ferramenta *Create Hexagon Tesselation***

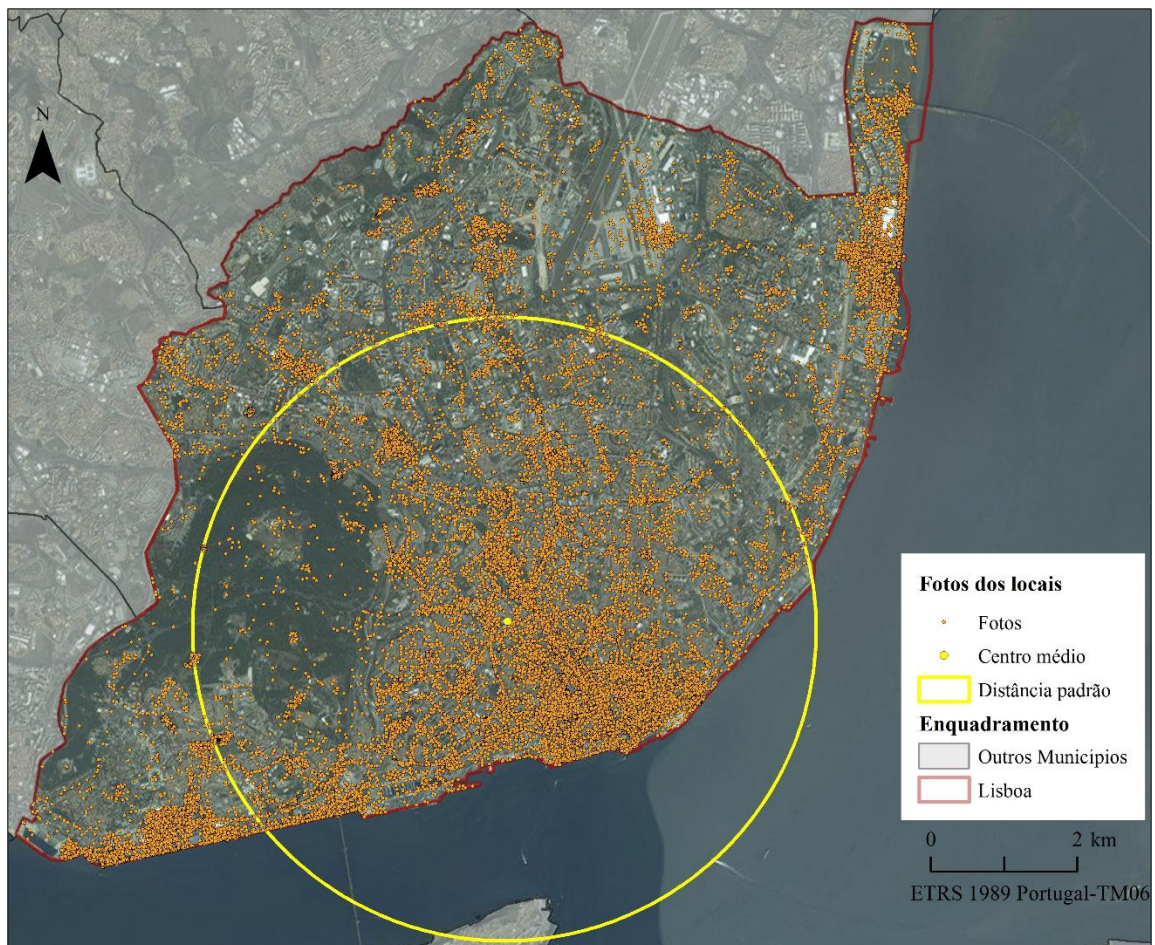




**Anexo 4. Características da distribuição espacial das fotografias dos visitantes, Centro médio e distância padrão.**



**Anexo 5. Características da distribuição espacial das fotografias dos locais, Centro médio e distância padrão.**



**Anexo 6. Fotografias dos locais e visitantes, segmentadas por meses e por anos.**

Anos	Meses													Média
	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total	
2005	-	-	-	-	-	-	-	-	-	1	2	-	3	-
2006	1	2	-	-	1	5	48	5	6	17	26	356	467	47
2007	523	755	652	435	551	375	375	628	561	606	472	496	6.429	536
2008	733	587	738	515	666	515	717	478	340	525	948	655	7.417	618
2009	607	839	785	844	881	517	509	671	667	381	604	1.220	8.525	710
2010	334	413	549	515	446	428	578	766	838	1.152	1.996	1.183	9.198	767
2011	1.061	608	822	730	913	1.214	1.076	874	757	900	893	885	10.733	894
2012	1.828	963	1.122	972	958	1.499	1.131	1.114	952	1.195	725	927	13.386	1.116
2013	1.048	786	696	849	974	840	919	997	1.743	1.526	1.257	1.314	12.949	1.079
2014	604	787	907	503	347	582	1.295	980	-	-	-	-	6.005	-
<b>Total</b>	6.739	5.740	6.271	5.363	5.737	5.975	6.648	6.513	5.864	6.303	6.923	7.036	75.112	
<b>Média</b>	876	707	766	694	770	770	758	790	837	898	985	954		

**Anexo 7. Número de utilizadores que carregaram fotos (considerados como visitantes e locais), segmentados por anos.**

	Anos											
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total	Média
<b>Nº Fotos</b>	3	467	6.429	7.417	8.525	9.198	10.733	13.386	12.949	6.005	75.112	9.805,3
<b>Nº Utilizadores</b>	2	148	1.575	1.465	1.252	1.033	1.138	901	720	355	8.589	1.154,9
<b>Ratio: Fotos / utilizador</b>	1,5	3,2	4,1	5,1	6,8	8,9	9,4	14,9	18,0	16,9	8,7	-
<b>% Fotos</b>	0	0,6	8,6	9,9	11,3	12,2	14,3	17,8	17,2	8,0	100	-
<b>% Utilizadores</b>	0	1,7	18,3	17,1	14,6	12,0	13,2	10,5	8,4	4,1	100	-

**Anexo 8. Amostra de fotografias dos visitantes, segmentadas por meses e por anos.**

Anos	Meses													Média
	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total	
2006	1				1	2	46	5	5	8	20	170	258	-
2007	277	501	393	216	170	209	205	290	216	214	99	198	2.988	249
2008	271	148	276	208	284	226	206	230	112	144	226	159	2.490	208
2009	182	303	242	294	221	167	137	171	115	102	205	387	2.526	211
2010	71	98	151	163	120	117	168	258	221	303	167	170	2.007	167
2011	299	111	216	177	155	182	160	75	141	224	189	125	2.054	171
2012	116	93	246	168	233	421	515	129	118	317	123	96	2.575	215
2013	151	90	55	253	92	192	184	213	273	60	147	134	1.844	154
2014	91	226	202	76	60	44	92	71					862	-
<b>Total</b>	1.459	1.570	1.781	1.555	1.336	1.560	1.713	1.442	1.201	1.372	1.176	1.439	17.604	

<b>Média</b>	195	192	226	211	182	216	225	195	171	195	165	181	
<b>%</b>	8,3	8,9	10,1	8,8	7,6	8,9	9,7	8,2	6,8	7,8	6,7	8,2	

**Anexo 9. Número de utilizadores considerados como visitantes, segmentadas por meses e por anos.**

Anos	Meses												Total	Média
	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez		
2006	1	-	-	-	1	2	2	3	3	6	7	85	110	-
2007	117	108	90	87	83	103	102	138	89	90	63	70	1.140	95,0
2008	85	63	88	69	109	81	85	90	59	66	87	71	953	79,4
2009	72	90	95	79	63	68	57	71	61	45	36	37	774	64,5
2010	34	37	40	38	49	40	38	67	42	68	59	58	570	47,5
2011	50	52	80	51	52	48	53	40	40	46	49	39	600	50,0
2012	42	31	27	47	45	34	38	35	41	38	33	26	437	36,4
2013	28	24	22	31	38	37	26	29	45	26	16	17	339	28,3
2014	27	23	22	22	21	22	22	10	-	-	-	-	169	-
<b>Total</b>	456	428	464	424	461	435	423	483	380	385	350	403	5.092*	

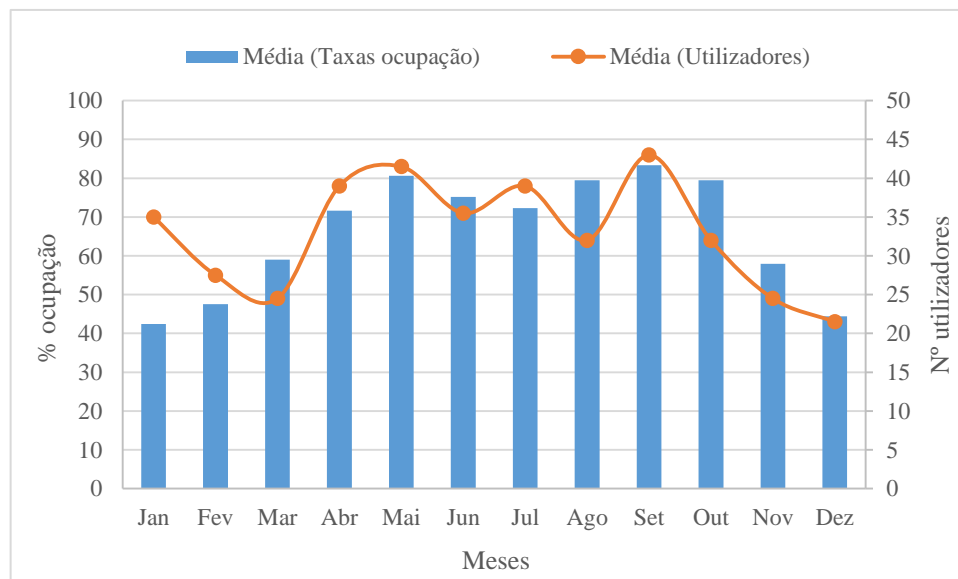
<b>Média</b>	61,1	57,9	63,1	57,4	62,7	58,7	57,0	67,1	53,9	54,1	49,0	45,4
<b>%</b>	9,0	8,4	9,1	8,3	9,1	8,5	8,3	9,5	7,5	7,6	6,9	7,9

\*Neste valor denota-se o acrescento de mas 10 utilizadores (visitantes) no total, em comparação com a síntese do número de utilizadores (visitantes) feita por anos (Figura 5.14 e Anexo 11), devido a que existem 10 utilizadores que tinham carregado fotos no final de um mês e no início do mês seguinte, dentro do mesmo ano. Como apresentado na secção 4.4.1.2. o total de utilizadores considerados como visitantes é de 5.082.



**Anexo 10. Taxas de ocupação hoteleira na cidade de Lisboa dos hotéis de 3, 4 e 5 estrelas para os anos 2012 e 2013; e número de utilizadores que carregaram fotos na rede Panoramio.**

	Taxas de ocupação			Utilizadores
	2012	2013	Média 2012-2013	Média 2012-2013
<b>Jan</b>	43,97	40,88	42,4	35,0
<b>Fev</b>	45,9	49,16	47,5	27,5
<b>Mar</b>	57,69	60,29	59,0	24,5
<b>Abr</b>	72,79	70,49	71,6	39,0
<b>Mai</b>	79,31	82,02	80,7	41,5
<b>Jun</b>	72,56	77,75	75,2	35,5
<b>Jul</b>	71,75	72,9	72,3	39,0
<b>Ago</b>	79,29	79,59	79,4	32,0
<b>Set</b>	82,76	84,03	83,4	43,0
<b>Out</b>	78,21	80,67	79,4	32,0
<b>Nov</b>	56,8	59,07	57,9	24,5
<b>Dez</b>	43,44	45,39	44,4	21,5



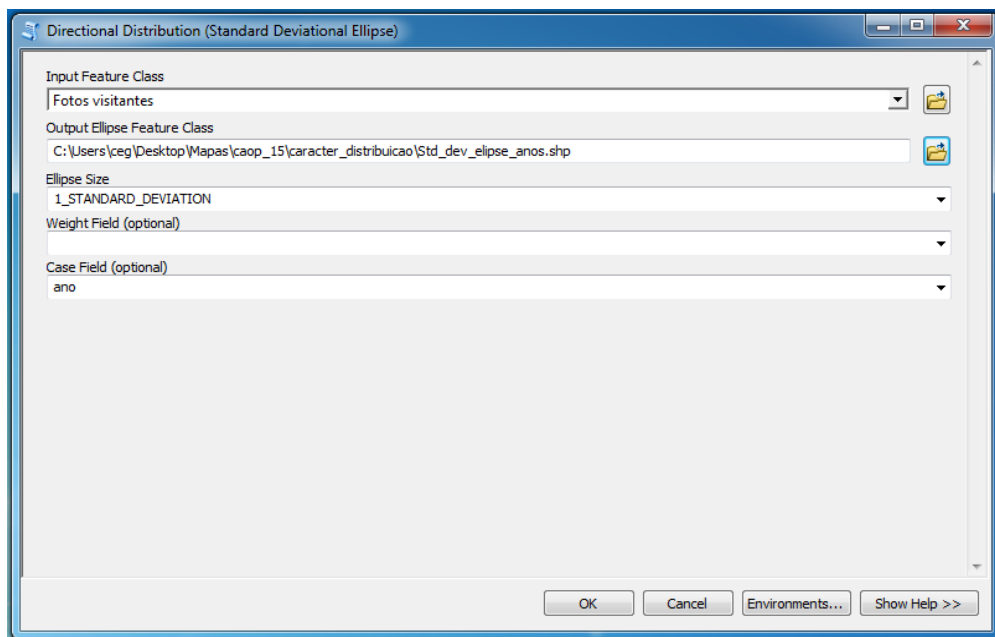
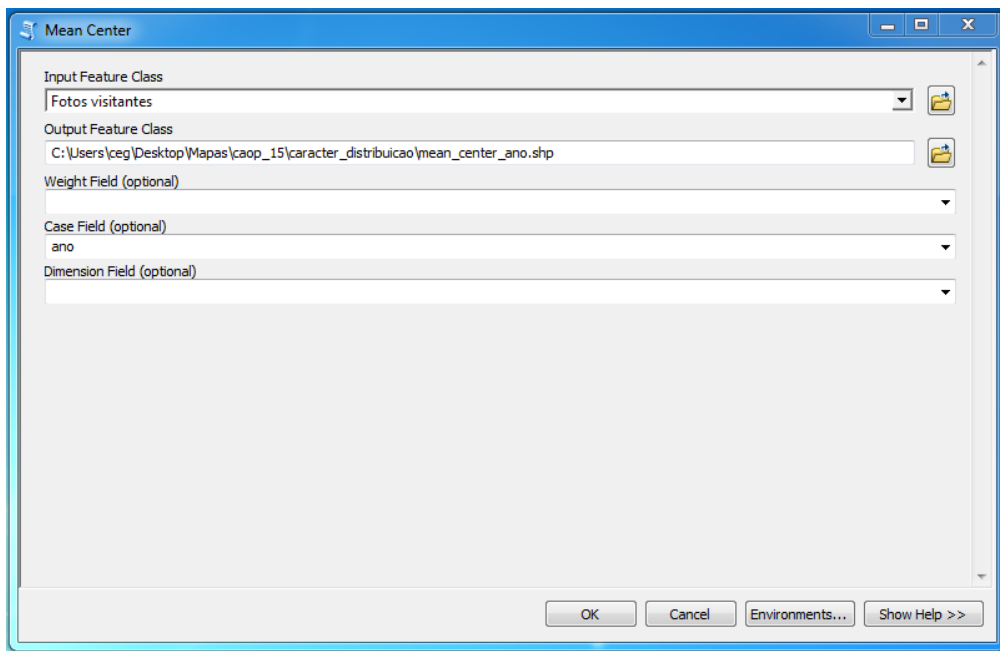
**Anexo 11. Número de visitantes e número de fotografias dos visitantes, segmentados por anos.**

	Anos										
	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total	Média
<b>Nº fotos</b>	258	2.988	2.490	2.526	2.007	2.054	2.575	1.844	862	17.604	2.354,9
<b>Nº Utilizadores</b>	110	1140	953	774	570	600	437	339	169	5.082	687,6
<b>Ratio Fotos utilizador /</b>											-
	2,3	2,6	2,6	3,3	3,5	3,4	5,9	5,4	5,1	3,5	
<b>% Fotos</b>	1,5	17,0	14,1	14,3	11,4	11,7	14,6	10,5	4,9	100	-
<b>% Utilizadores</b>	2,2	22,4	18,8	15,2	11,2	11,8	8,6	6,7	3,3	100	-

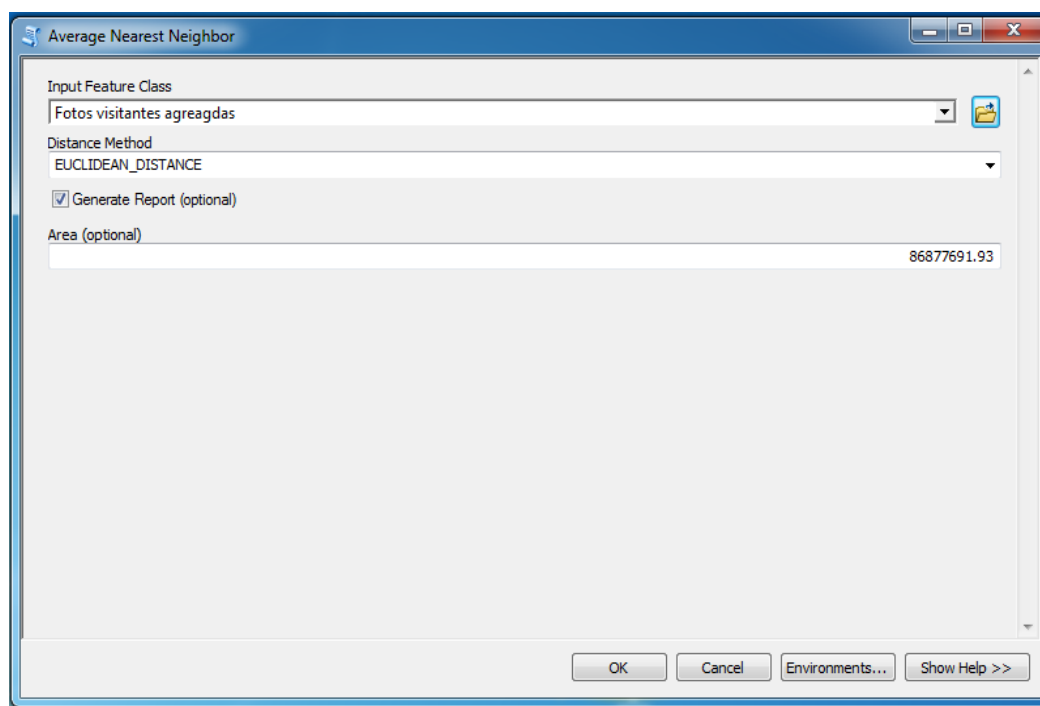
**Anexo 12. Fotos por Km<sup>2</sup>, segmentadas por freguesias.**

Nº	Freguesia	Nº Fotos	Área (km2)	Fotos/área
1	Santa Maria Maior	4.242	1,51	2.815,48
2	Misericórdia	1.312	1,11	1.177,59
3	Belém	3.820	5,72	6.67,92
4	São Vicente	614	1,25	490,15
5	Parque das Nações	2.301	4,84	475,69
6	Santo António	705	1,49	471,78
7	Avenidas Novas	974	2,99	325,27
8	Estrela	640	2,83	226,18
9	São Domingos de Benfica	690	4,29	160,68
10	Arroios	273	2,13	128,30
11	Alcântara	398	4,43	89,83
12	Campolide	221	2,77	79,66
13	Campo de Ourique	126	1,65	76,30
14	Areeiro	120	1,74	68,86
15	Penha de França	118	2,20	53,52
16	Alvalade	197	5,34	36,88
17	Ajuda	100	2,88	34,76
18	Lumiar	191	6,57	29,05
19	Olivais	234	8,09	28,93
20	Benfica	127	8,02	15,83
21	Marvila	98	6,27	15,63
22	Carnide	55	3,69	14,91
23	Beato	22	1,68	13,08
24	Santa Clara	26	3,36	7,75
Total		17.604,00	86,88	7.504,03
<b>Média</b>		312,67		

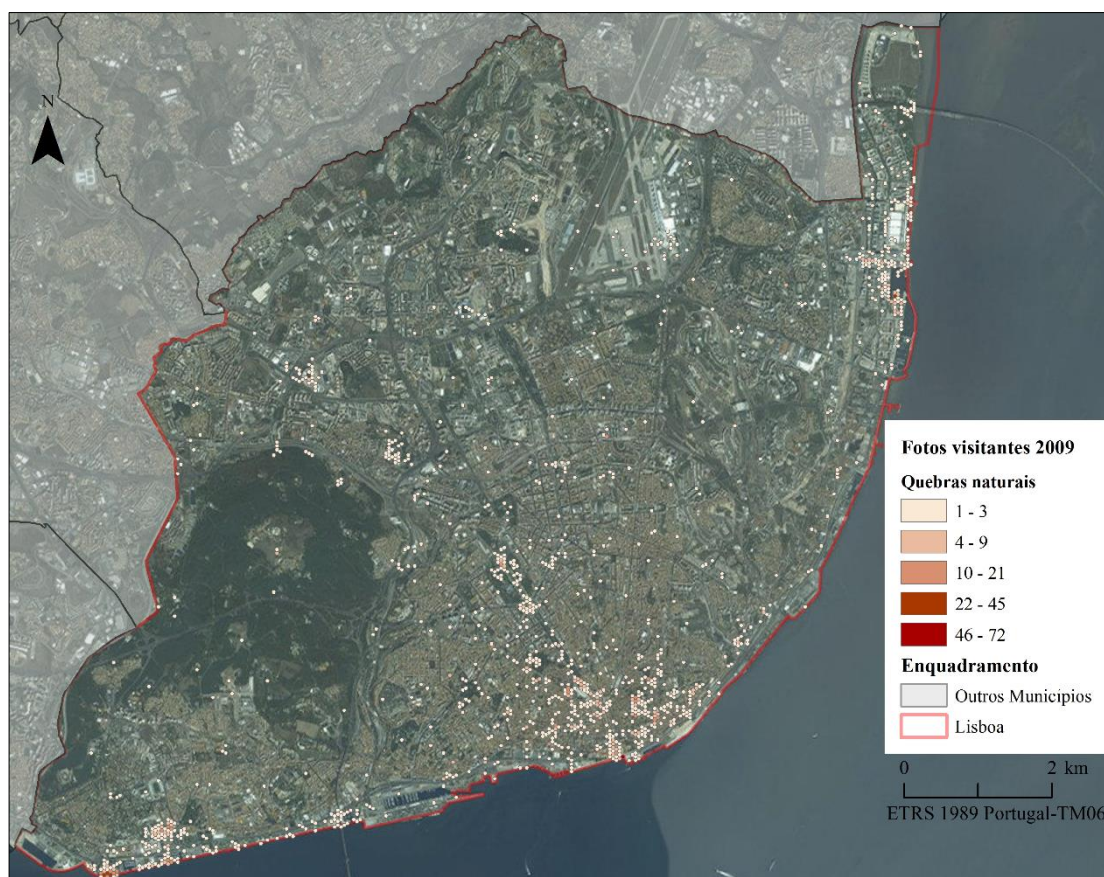
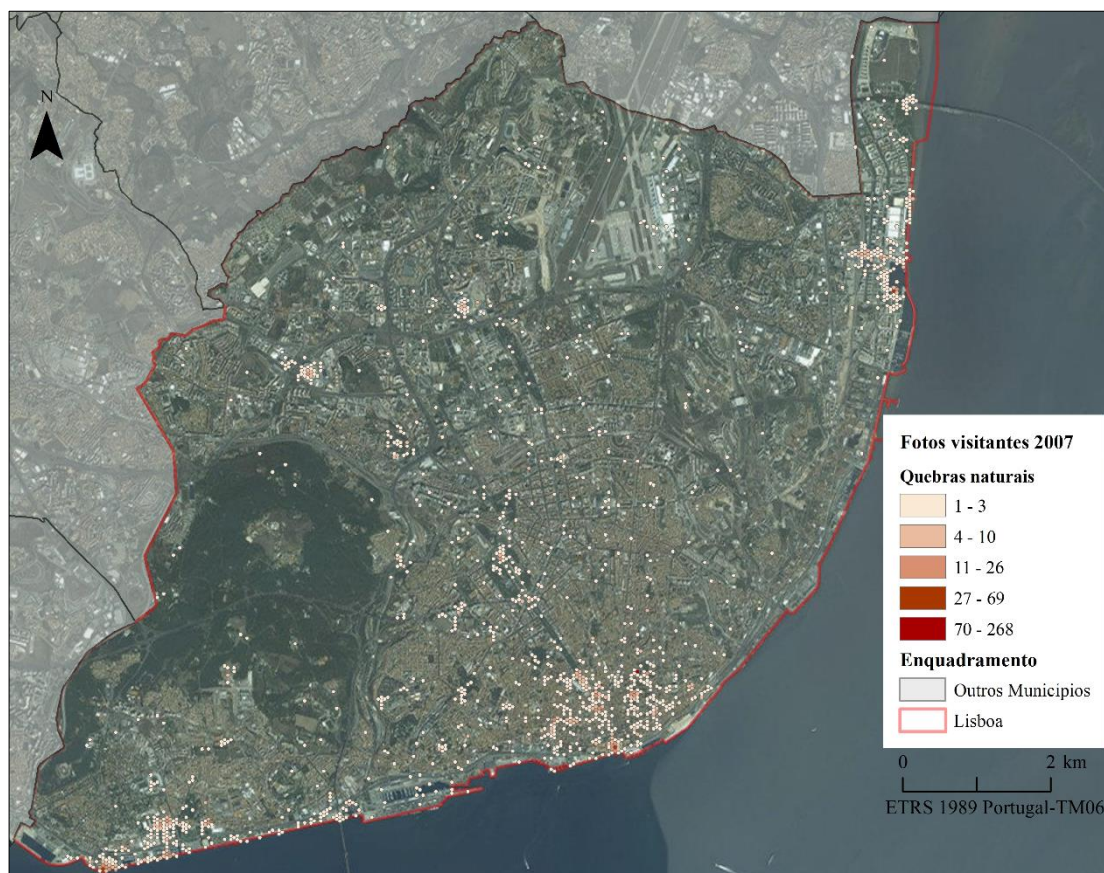
### Anexo 13. Parametrização das ferramentas: Centro médio e Elipse do desvio padrão.



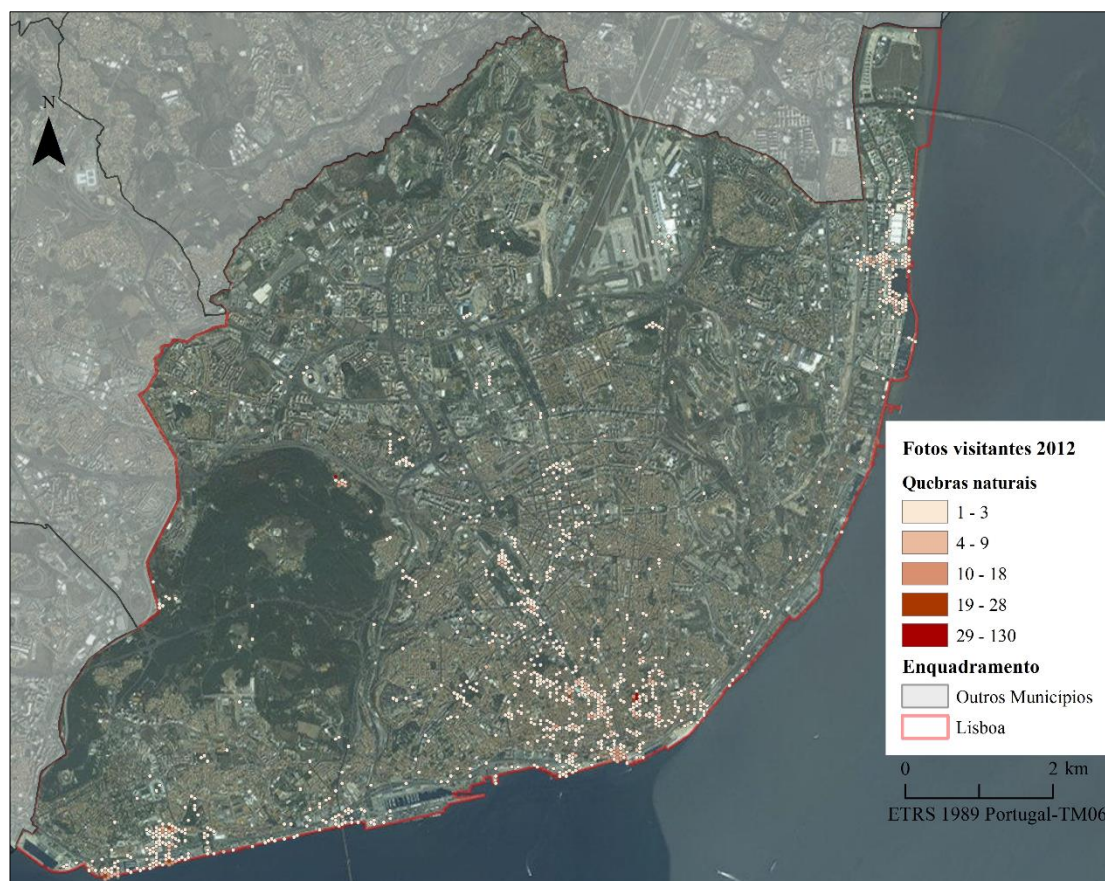
#### Anexo 14. Parametrização da ferramenta: *Nearest Neighbor Index*.



**Anexo 15. Fotografias dos visitantes do ano 2007, 2009 e 2012 agregadas por unidades hexagonais.**

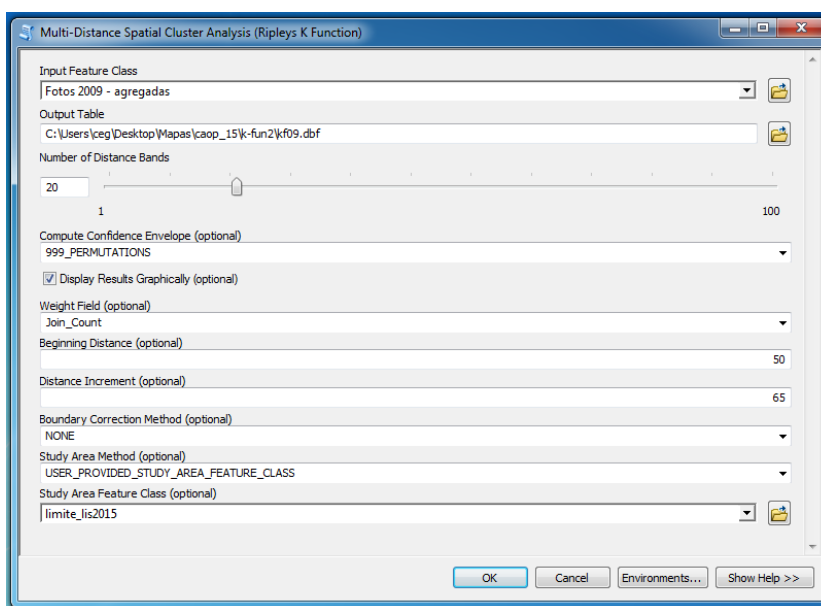
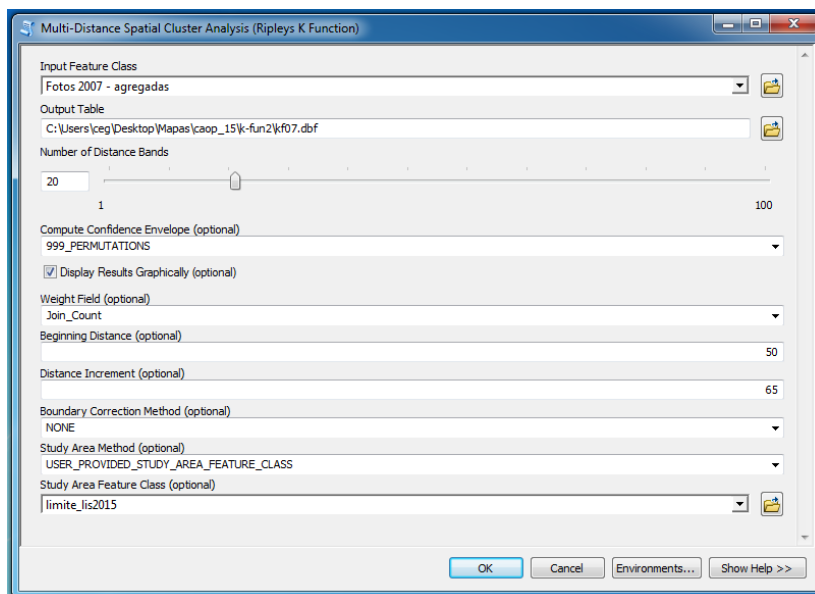


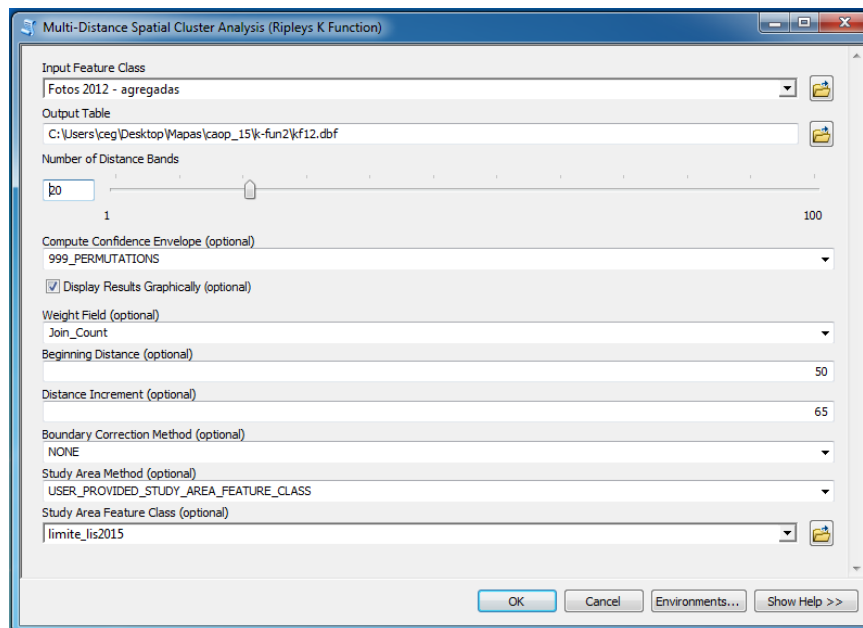




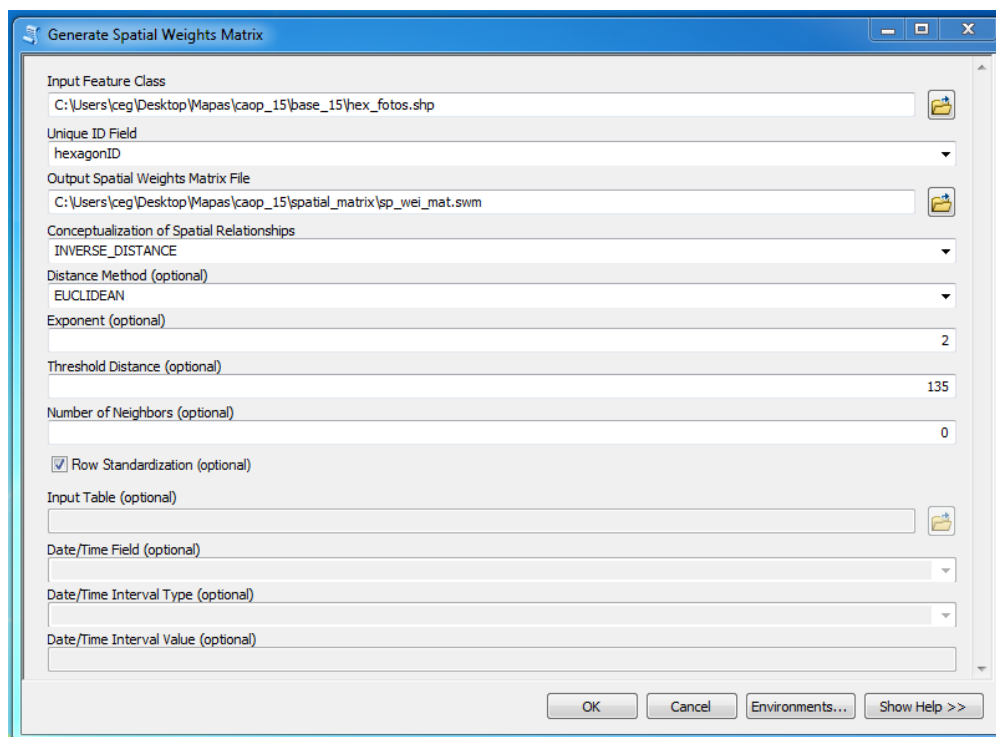


**Anexo 16. Parametrização da ferramenta: *K-Function (Multi-Distance Spatial Cluster Analysis)* para os anos 2007, 2009 e 2012.**

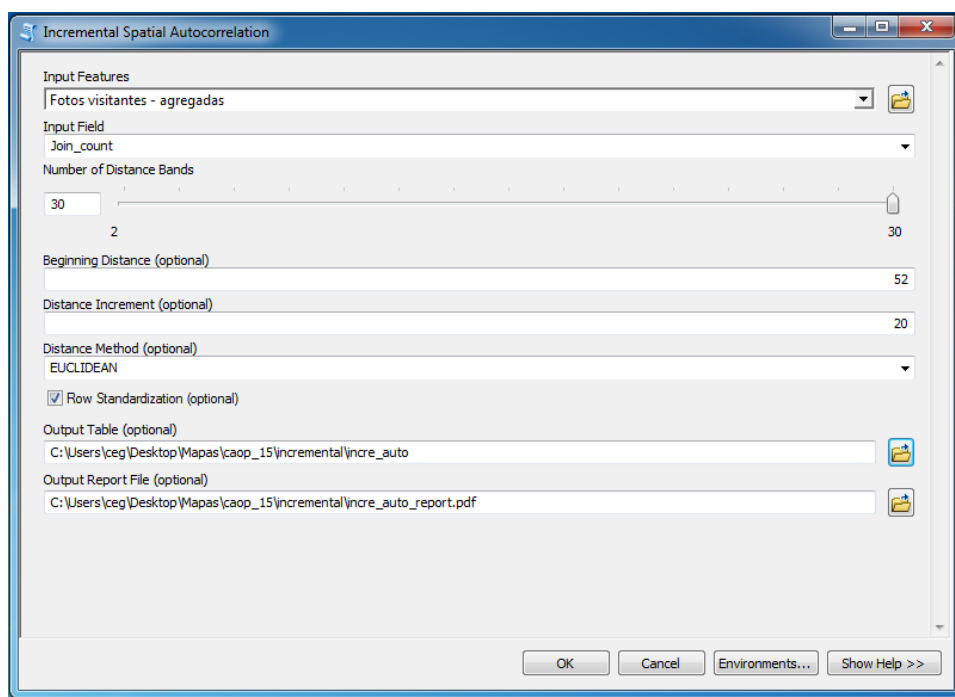




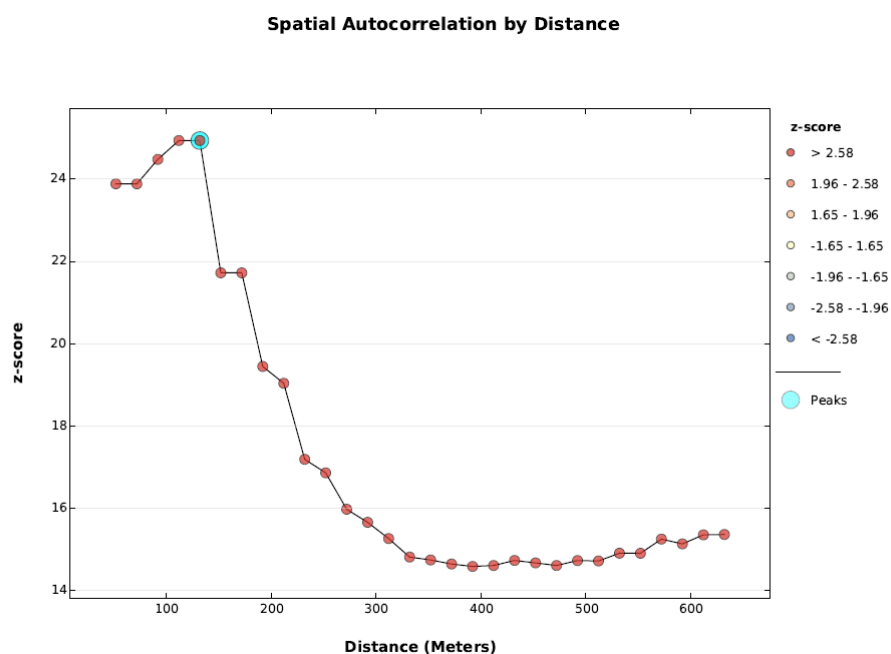
## Anexo 17. Parametrização da ferramenta: *Spatial Weight Matrix*.



## Anexo 18. Parametrização da ferramenta: *Spatial Incremental Autocorrelation*.



## Anexo 19. Output da ferramenta: *Spatial Incremental Autocorrelation*.



### Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
52.00*	0.355758	-0.000291	0.000222	23.885049	0.000000
72.00*	0.355758	-0.000291	0.000222	23.885049	0.000000
92.00	0.292624	-0.000270	0.000143	24.480515	0.000000
112.00	0.259589	-0.000270	0.000109	24.938667	0.000000
132.00	0.259589	-0.000270	0.000109	24.938667	0.000000
152.00	0.180756	-0.000270	0.000069	21.721266	0.000000
172.00	0.180756	-0.000270	0.000069	21.721266	0.000000
192.00	0.141094	-0.000270	0.000053	19.444029	0.000000
212.00	0.133422	-0.000270	0.000049	19.034517	0.000000
232.00	0.108049	-0.000270	0.000040	17.185559	0.000000
252.00	0.103126	-0.000270	0.000038	16.856930	0.000000
272.00	0.092274	-0.000270	0.000034	15.971613	0.000000
292.00	0.087063	-0.000270	0.000031	15.652898	0.000000
312.00	0.080402	-0.000270	0.000028	15.258750	0.000000
332.00	0.073686	-0.000270	0.000025	14.808681	0.000000
352.00	0.068988	-0.000270	0.000022	14.737867	0.000000
372.00	0.066731	-0.000270	0.000021	14.639534	0.000000
392.00	0.063111	-0.000270	0.000019	14.580784	0.000000
412.00	0.059687	-0.000270	0.000017	14.603461	0.000000
432.00	0.058929	-0.000270	0.000016	14.726377	0.000000
452.00	0.055563	-0.000270	0.000014	14.666692	0.000000
472.00	0.054527	-0.000270	0.000014	14.604065	0.000000
492.00	0.052492	-0.000270	0.000013	14.723636	0.000000
512.00	0.050345	-0.000270	0.000012	14.713724	0.000000
532.00	0.048499	-0.000270	0.000011	14.902737	0.000000
552.00	0.047375	-0.000270	0.000010	14.900389	0.000000
572.00	0.046386	-0.000270	0.000009	15.245278	0.000000
592.00	0.044629	-0.000270	0.000009	15.129023	0.000000
612.00	0.043787	-0.000270	0.000008	15.349177	0.000000
632.00	0.042650	-0.000270	0.000008	15.358443	0.000000

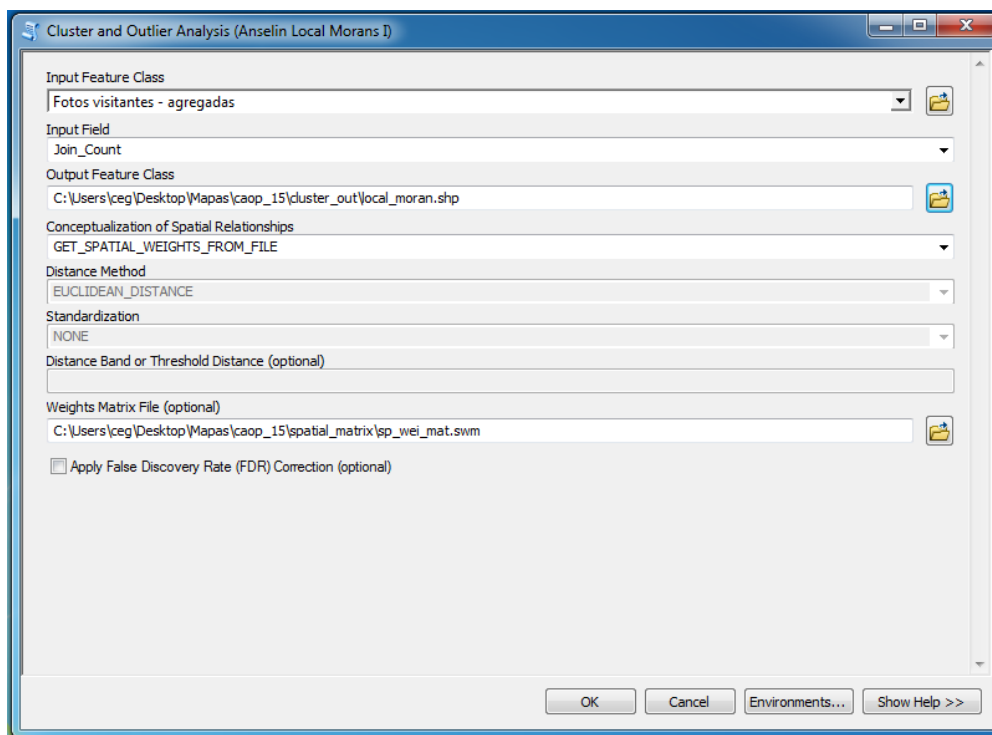
First Peak (Distance, Value): 132.00, 24.938667

Max Peak (Distance, Value): 132.00, 24.938667

Distance measured in Meters

\* At least one distance increment resulted in features with no neighbors which may invalidate the significance of the corresponding results.

## Anexo 20. Parametrização da ferramenta: Análise de *Cluster* e *Outliers*.



## Anexo 21. Parametrização da ferramenta: *Create space time cube*.

Geoprocessing

Create Space Time Cube

Parameters | Environments

Input Features  
sample\_fotos\_cb2.shp

Output Space Time Cube  
E:\tesis\Mapas\caop\_15\cubo\_3\arc\_pro\cubo\_1.nc

Time Field  
upload\_dat

Template Cube

Time Step Interval  
1 Months

Time Step Alignment  
End time

Distance Interval  
50 Meters

Summary Fields (+) (-)

Field  
Statistic  
Fill Empty Bins with

Aggregation Shape Type  
Hexagon grid

Run

Project | Geoprocessing | Symbology

## Anexo 22. Parametrização da ferramenta: *Emerging hot spot analysis*.

Geoprocessing

Emerging Hot Spot Analysis

Parameters | Environments

Input Space Time Cube  
E:\tesis\Mapas\caop\_15\cubo\_3\arc\_pro\cubo\_1.nc

Analysis Variable  
COUNT

Output Features  
cubo\_1\_EHotSpot

Neighborhood Distance  
135 Meters

Neighborhood Time Step  
11

Polygon Analysis Mask

Run

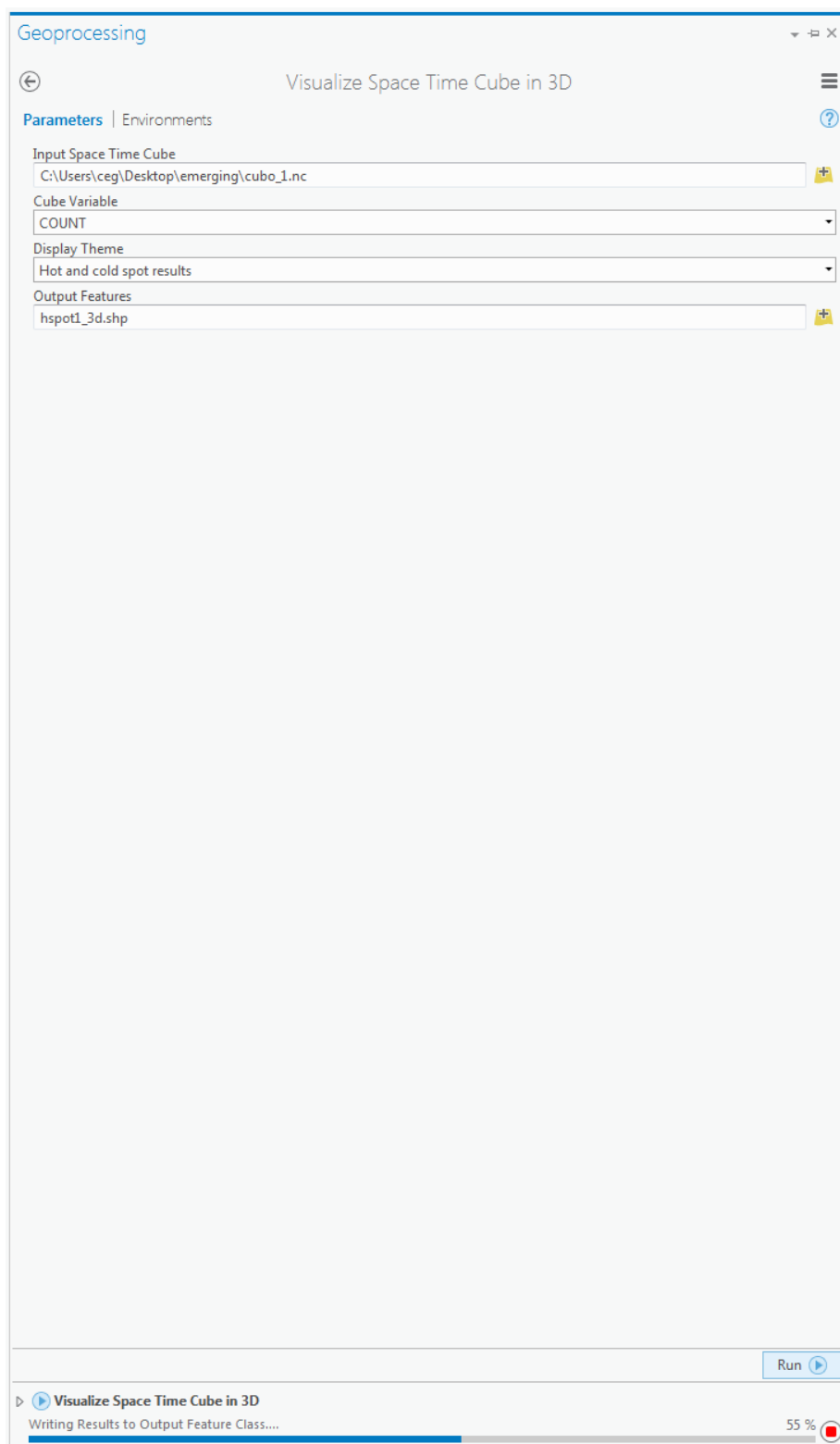
✓ Create Space Time Cube  
Completed successfully

Execute Tool

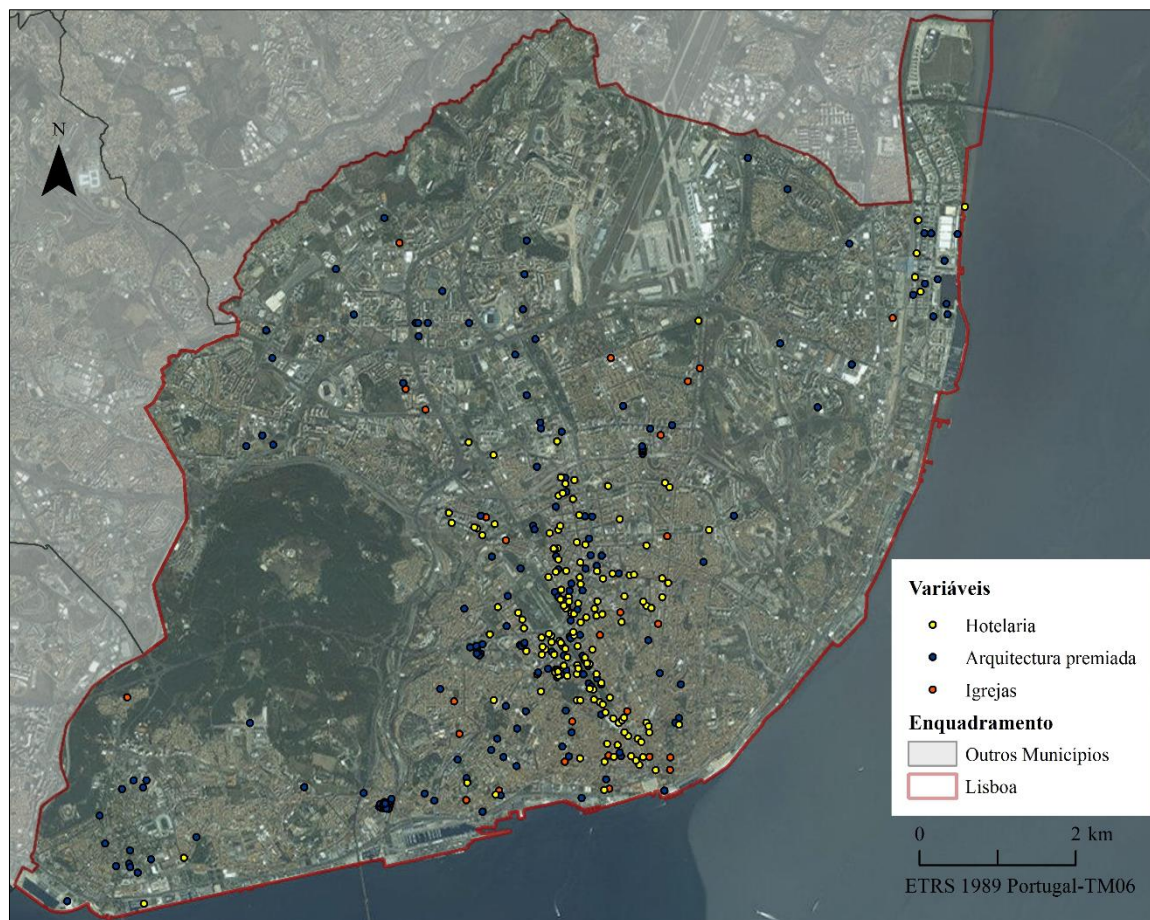
Project | Geoprocessing | Symbology



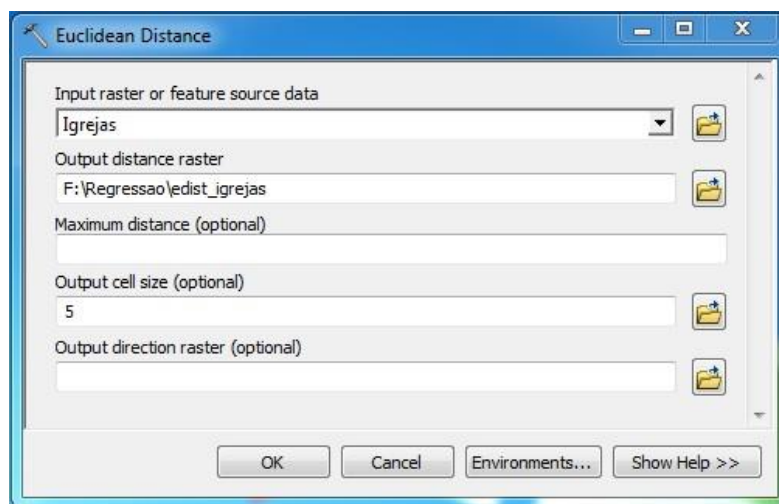
## Anexo 23. Parametrização da ferramenta: *Visualize space time cube in 3D*.



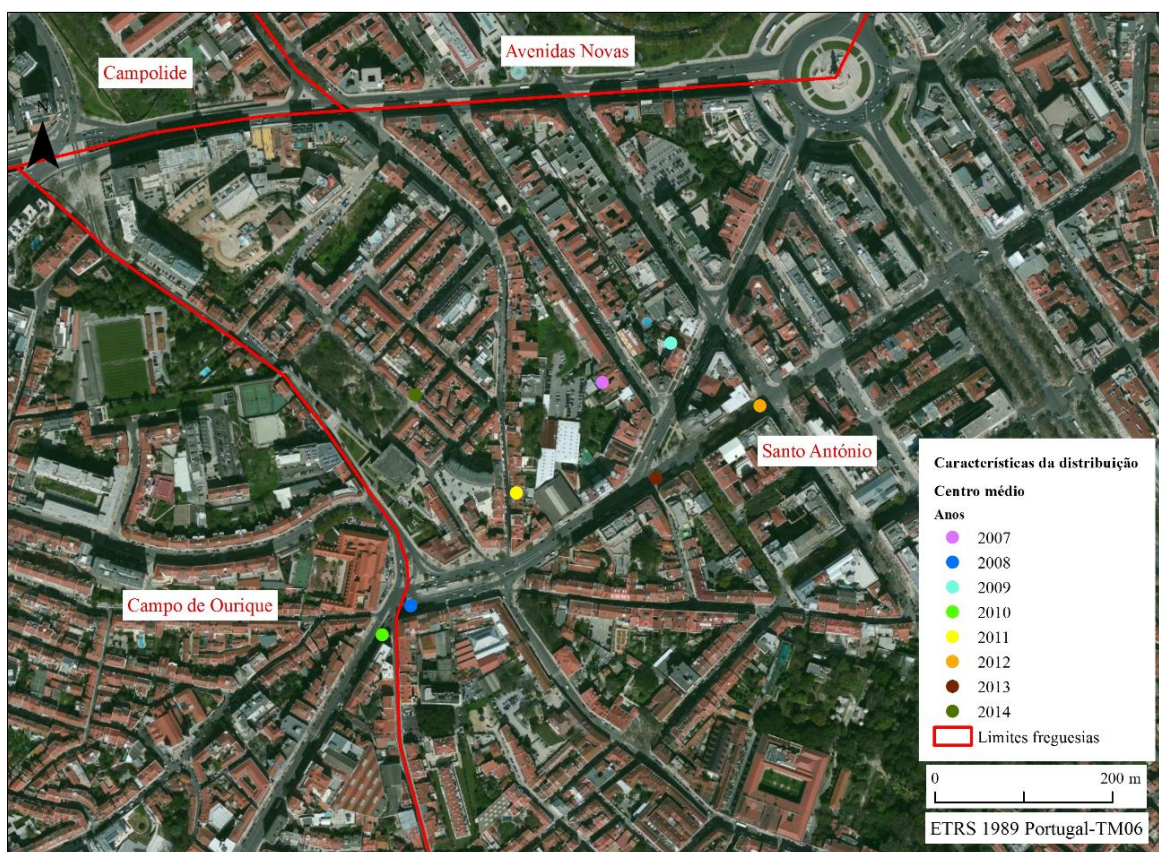
**Anexo 24. Localização das variáveis seleccionadas: Hotelaria, Arquitectura premiada, Igrejas.**



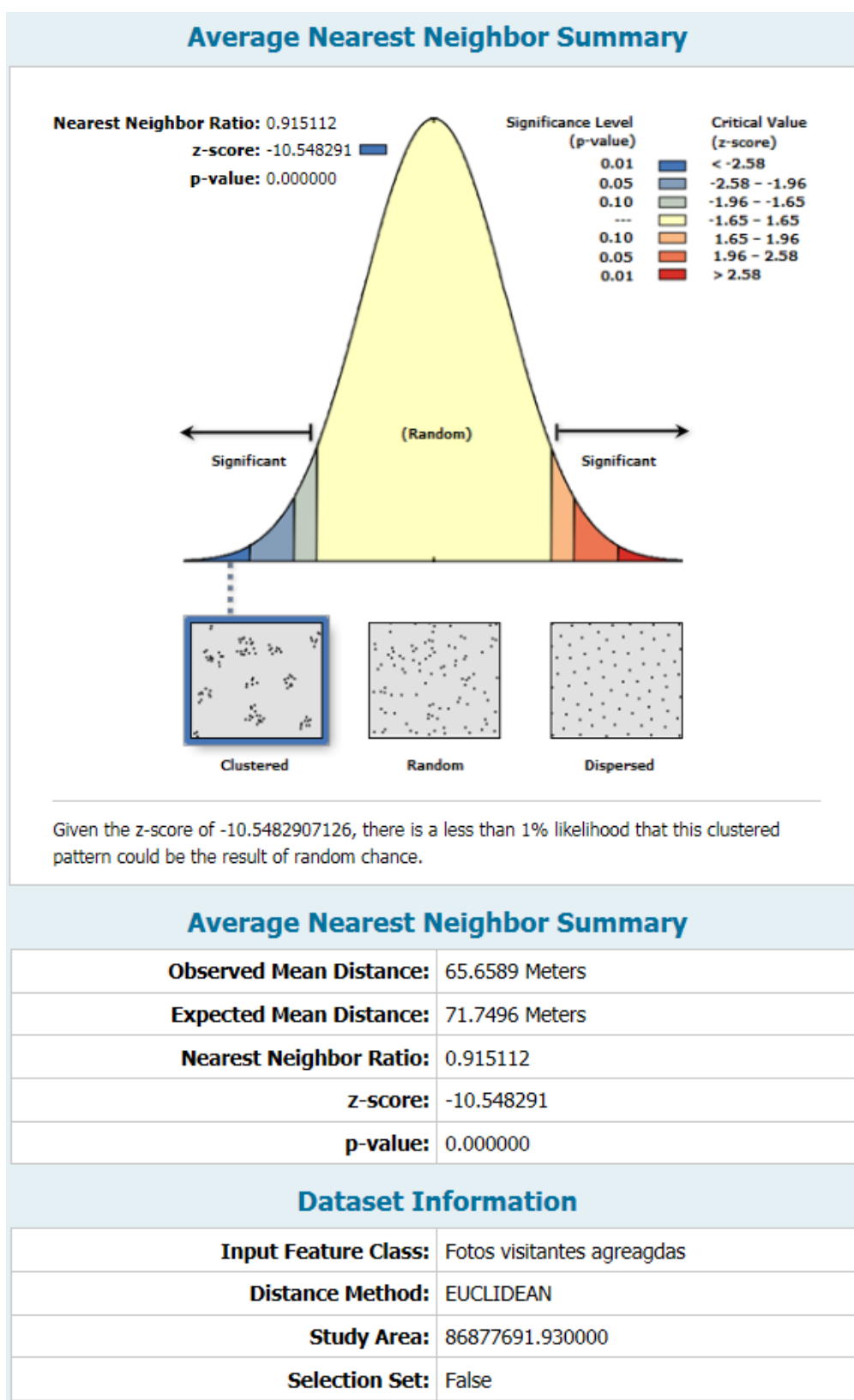
**Anexo 25. Parametrização da ferramenta *Euclidean distance*, realizada para a variável Igrejas.**



**Anexo 26. Centros médios dos 8 anos em análise (de 2007 a 2014).**



## Anexo 27. Relatório do método *Average nearest neighbor index*.

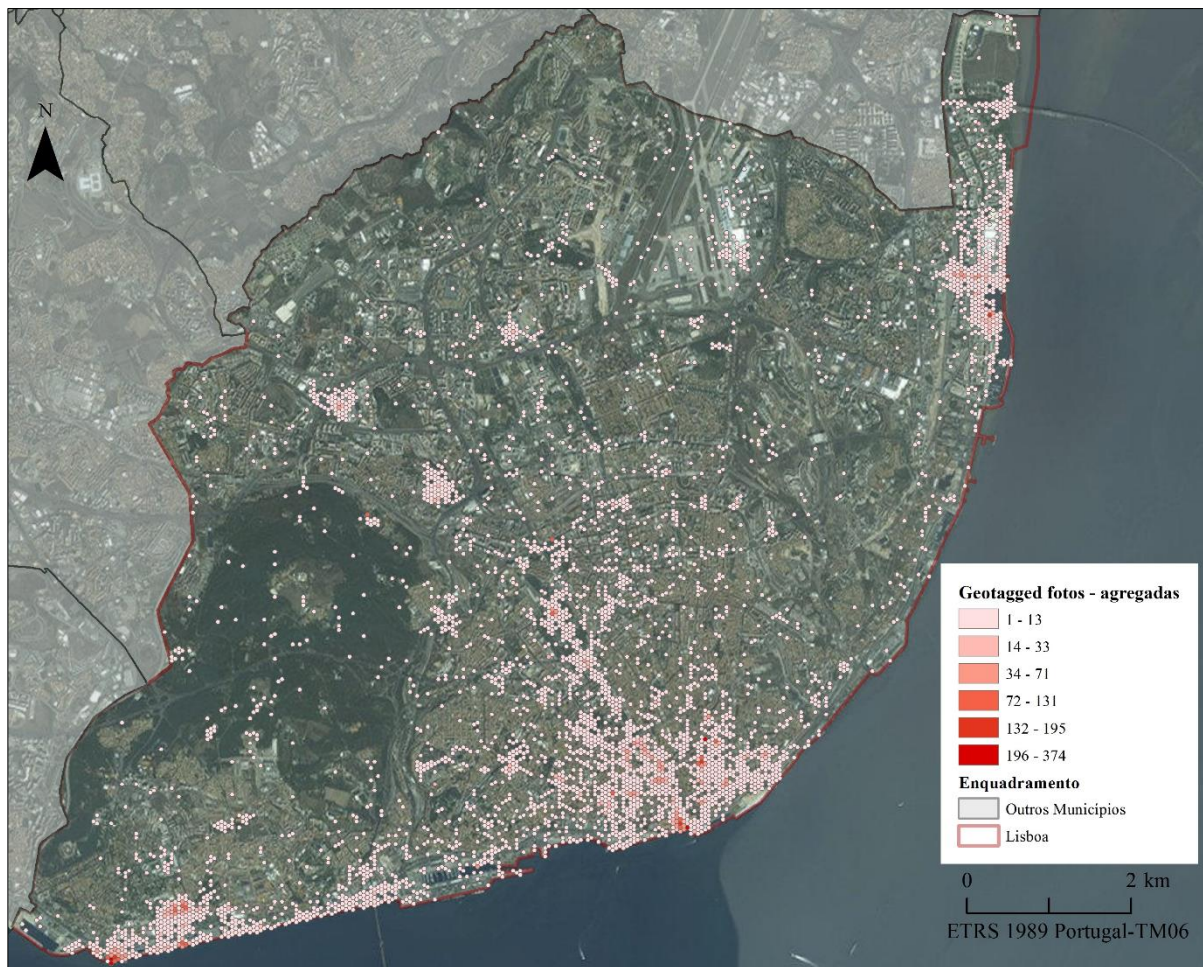




**Anexo 28. Resultados do método de *K-function*, valores esperados e observados da função e limites de confiança estatística.**

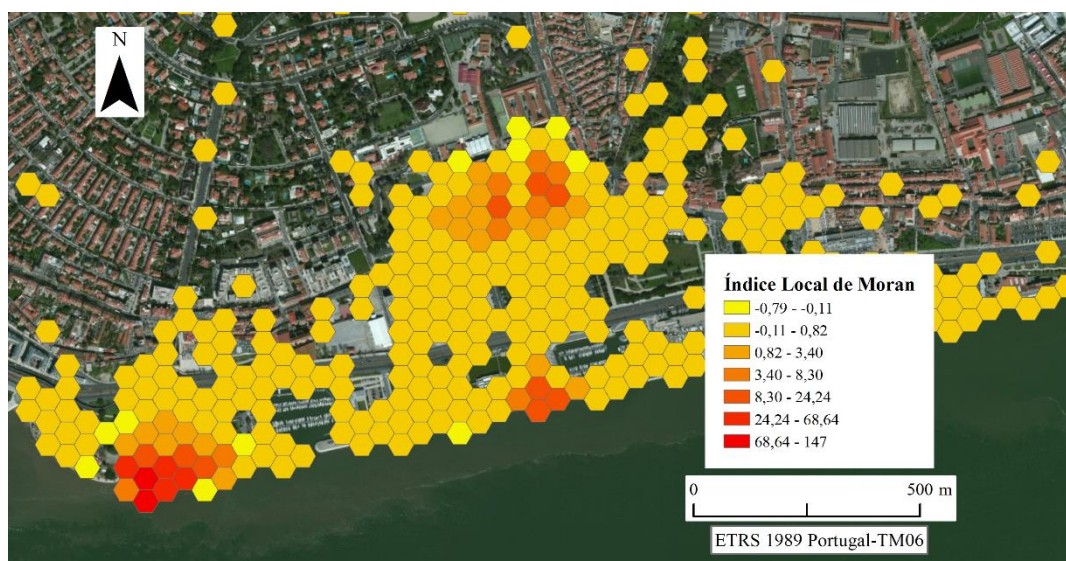
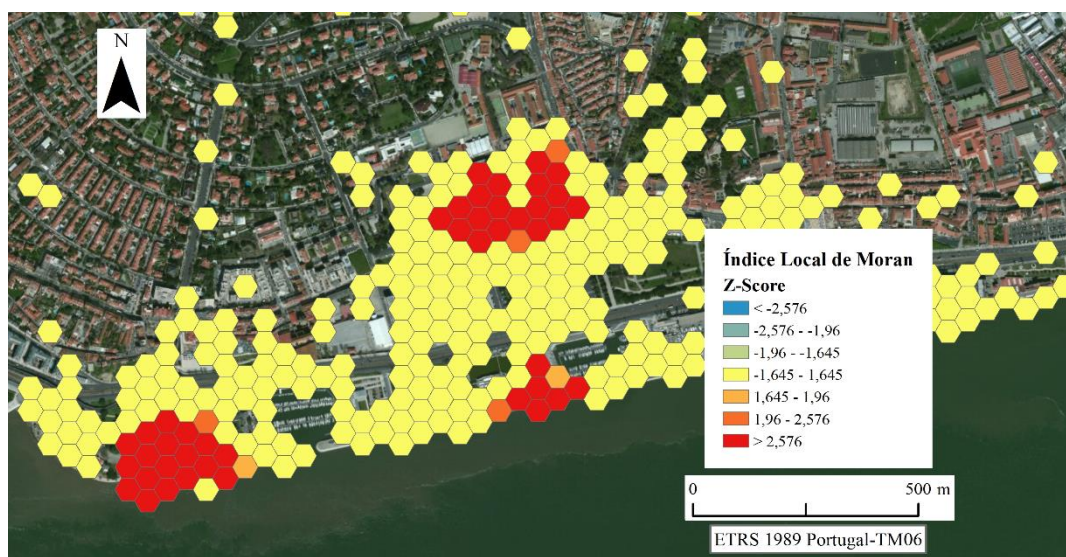
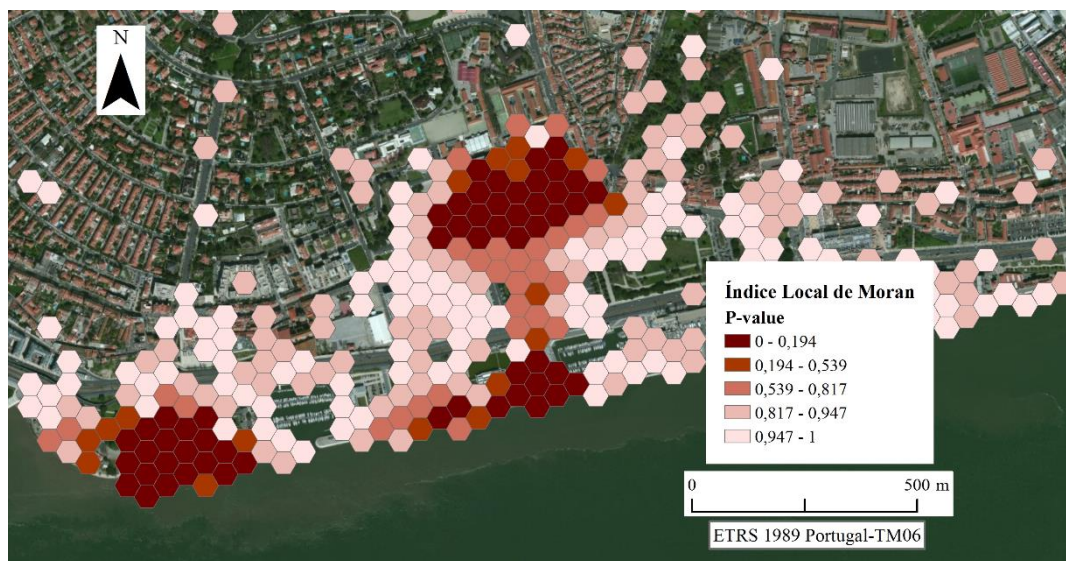
Valores esperados			Valores observados			Diferença entre valores			Limite de confiança baixo			Limite de confiança alto		
2007	2009	2012	2007	2009	2012	2007	2009	2012	2007	2009	2012	2007	2009	2012
50,0	50,0	50,0	363,0	334,6	437,0	313,0	284,6	387,0	140,2	158,5	153,8	296,0	222,6	337,0
115,0	115,0	115,0	550,5	551,9	617,8	435,5	436,9	502,8	257,8	287,1	289,4	467,4	369,4	428,9
180,0	180,0	180,0	686,3	696,7	743,6	506,3	516,7	563,6	373,4	410,4	403,0	652,7	511,9	567,8
245,0	245,0	245,0	821,3	835,2	878,8	576,3	590,2	633,8	510,9	538,2	524,3	810,0	648,6	759,1
310,0	310,0	310,0	960,6	944,4	1.011,6	650,6	634,4	701,6	608,9	648,5	655,1	915,8	769,9	878,2
375,0	375,0	375,0	1.077,9	1.047,2	1.103,0	702,9	672,2	728,0	699,6	739,9	757,7	999,5	881,6	980,0
440,0	440,0	440,0	1.185,3	1.159,1	1.227,6	745,3	719,1	787,6	809,1	836,3	874,4	1.093,6	1.015,5	1.083,2
505,0	505,0	505,0	1.283,1	1.257,1	1.356,0	778,1	752,1	851,0	888,3	921,3	974,4	1.179,1	1.130,7	1.217,3
570,0	570,0	570,0	1.406,4	1.356,8	1.480,9	836,4	786,8	910,9	972,8	1.008,5	1.079,0	1.292,5	1.240,3	1.354,5
635,0	635,0	635,0	1.516,5	1.446,2	1.576,9	881,5	811,2	941,9	1.046,7	1.084,2	1.155,8	1.411,4	1.335,2	1.452,4
700,0	700,0	700,0	1.628,5	1.554,9	1.691,9	928,5	854,9	991,9	1.133,1	1.170,6	1.243,8	1.508,6	1.439,9	1.583,9
765,0	765,0	765,0	1.751,4	1.648,1	1.791,1	986,4	883,1	1.026,1	1.201,3	1.242,7	1.320,6	1.596,4	1.532,4	1.684,8
830,0	830,0	830,0	1.860,5	1.751,1	1.901,7	1.030,5	921,1	1.071,7	1.270,0	1.313,2	1.401,1	1.665,0	1.623,5	1.784,9
895,0	895,0	895,0	1.949,9	1.843,7	2.016,6	1.054,9	948,7	1.121,6	1.342,9	1.379,7	1.476,9	1.750,9	1.700,0	1.861,6
960,0	960,0	960,0	2.031,4	1.929,5	2.121,9	1.071,4	969,5	1.161,9	1.409,7	1.449,1	1.548,0	1.840,6	1.775,0	1.951,6
1.025,0	1.025,0	1.025,0	2.109,1	1.997,9	2.189,7	1.084,1	972,9	1.164,7	1.463,1	1.512,2	1.605,3	1.903,8	1.838,8	2.032,7
1.090,0	1.090,0	1.090,0	2.201,2	2.071,1	2.260,1	1.111,2	981,1	1.170,1	1.521,2	1.585,8	1.671,5	1.971,7	1.908,4	2.155,3
1.155,0	1.155,0	1.155,0	2.273,4	2.134,7	2.324,5	1.118,4	979,7	1.169,5	1.574,6	1.645,3	1.732,9	2.040,2	1.971,7	2.234,1
1.220,0	1.220,0	1.220,0	2.311,1	2.182,6	2.379,5	1.091,1	962,6	1.159,5	1.623,1	1.712,4	1.796,7	2.095,3	2.023,7	2.305,7
1.285,0	1.285,0	1.285,0	2.349,4	2.227,5	2.428,7	1.064,4	942,5	1.143,7	1.668,0	1.768,1	1.851,1	2.140,8	2.076,8	2.368,9

**Anexo 29. Geotagged fotos dos visitantes agregadas por unidades hexagonais (classificadas segundo as quebras naturais).**



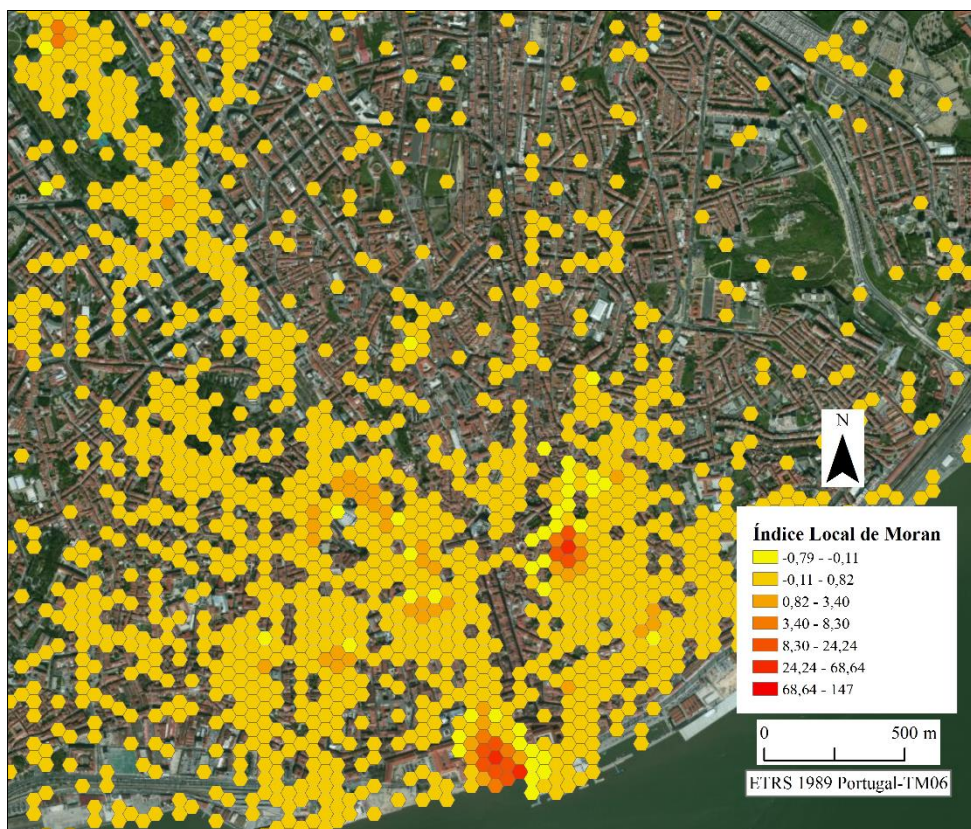
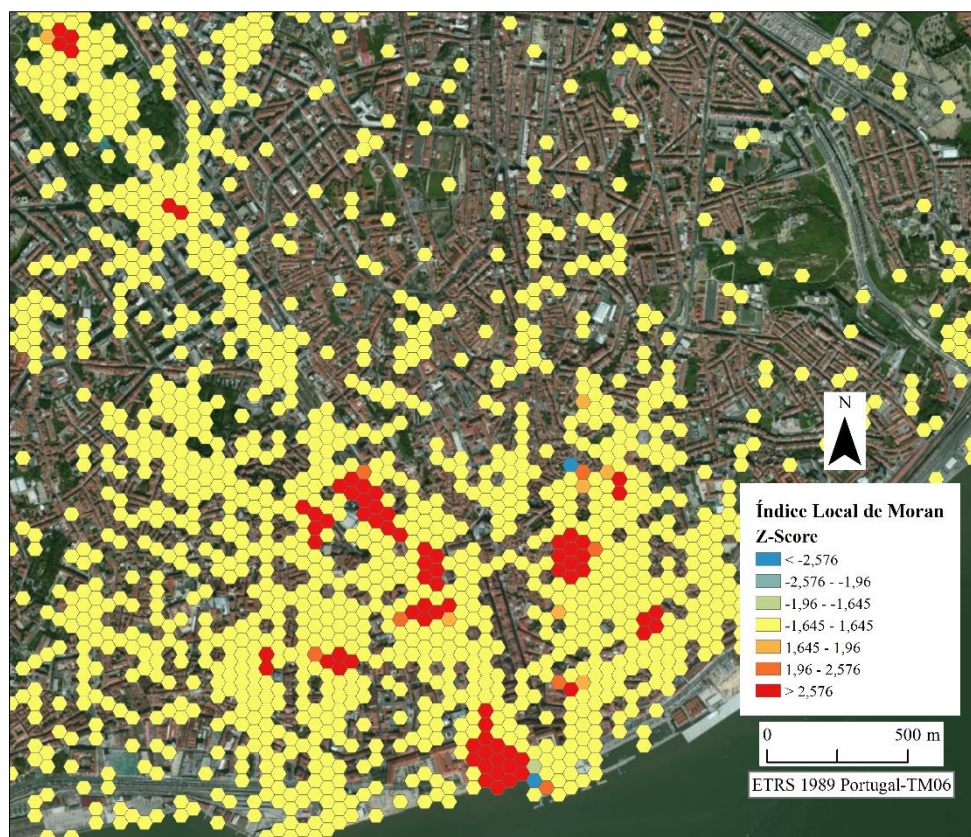


**Anexo 30. Análise de *clusters* e *outliers* das aglomerações localizadas na parte sul da cidade.**

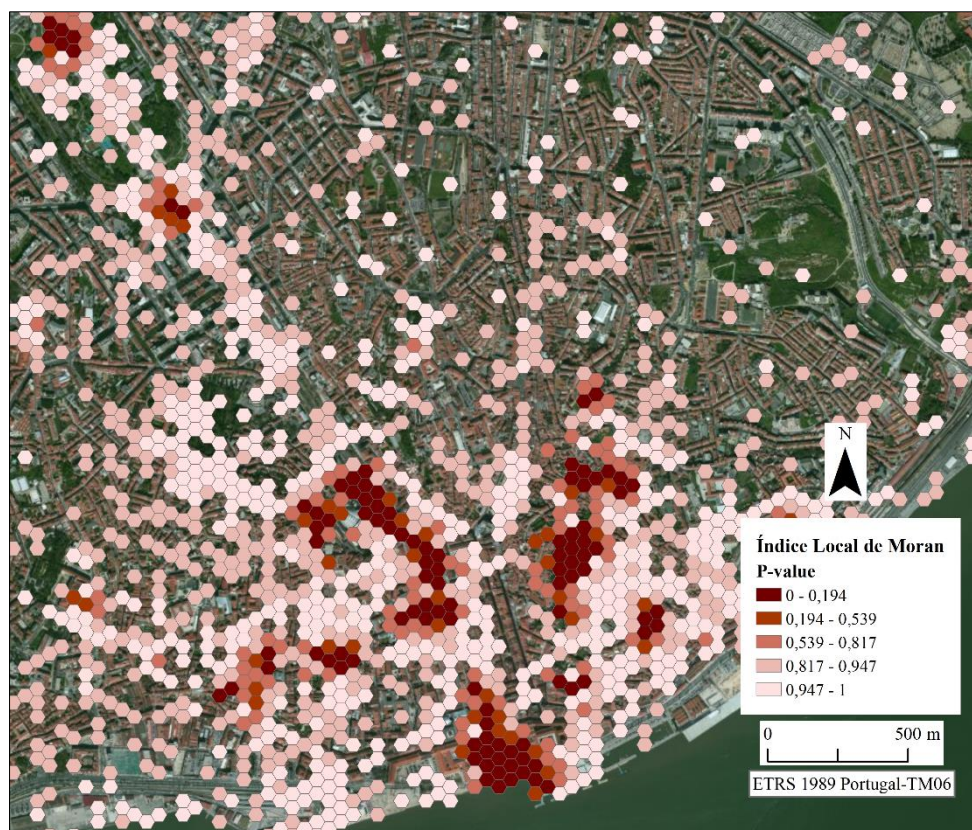




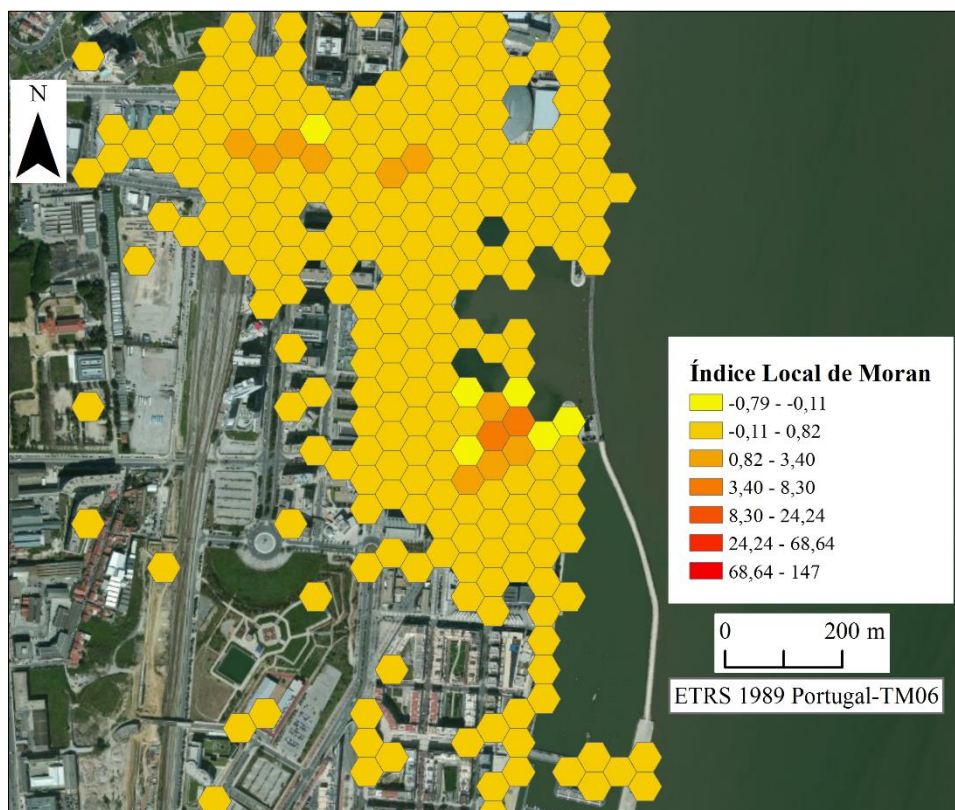
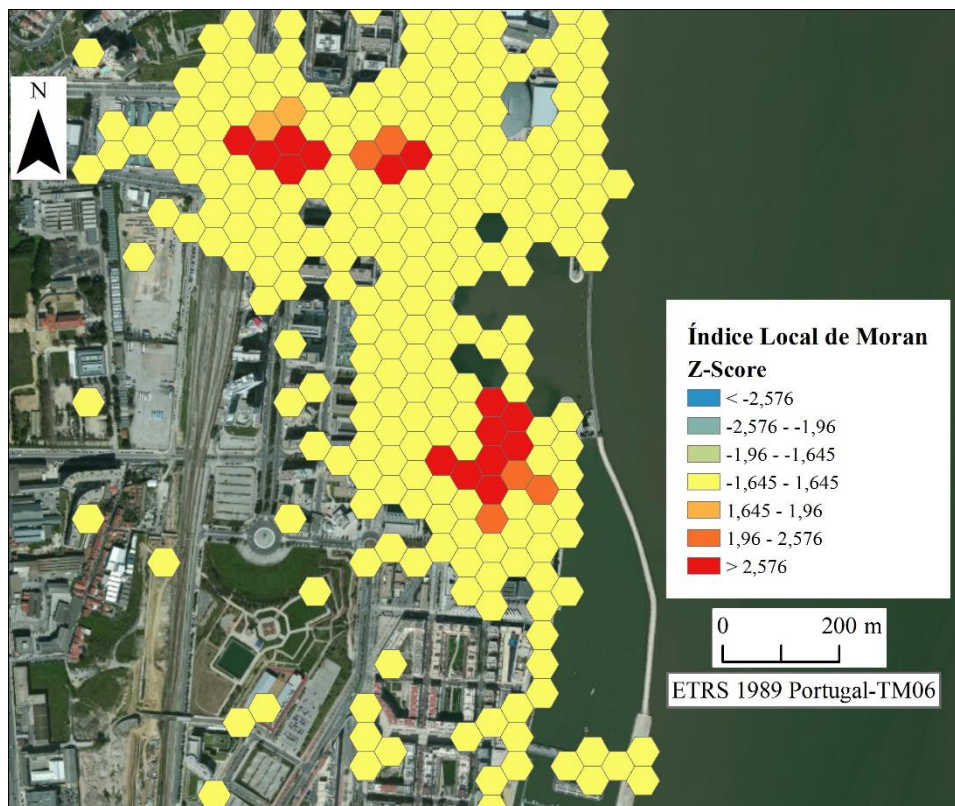
**Anexo 31. Análise de *clusters* e *outliers* das aglomerações localizadas no centro da cidade.**



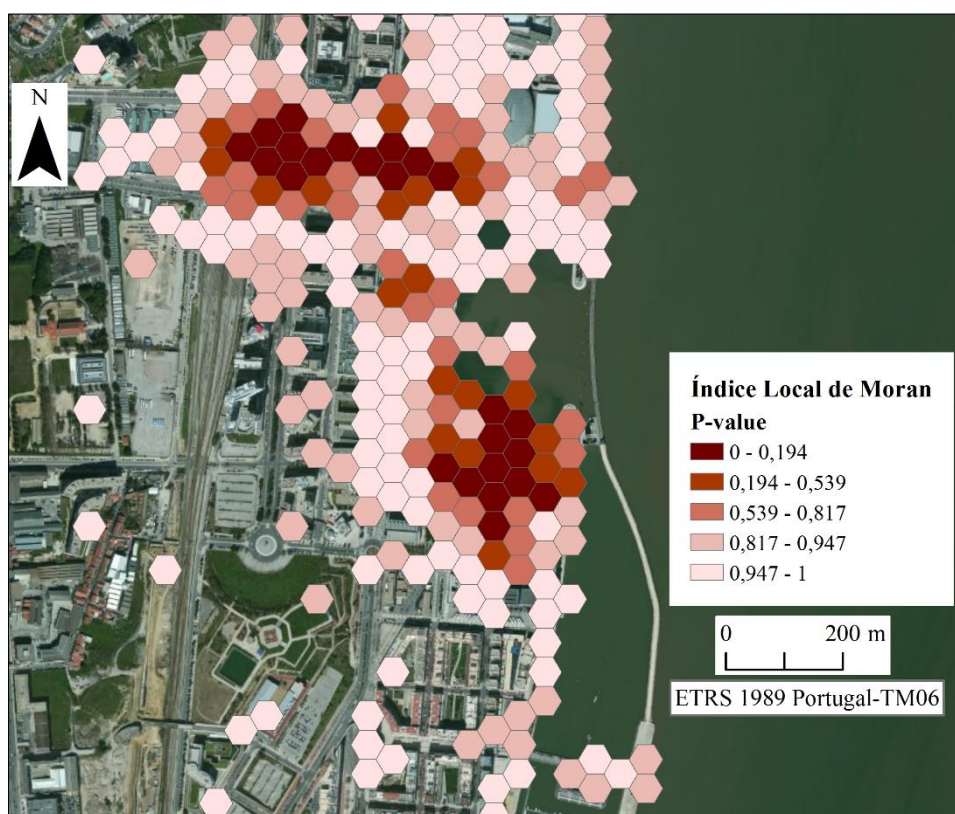




**Anexo 32. Análise de *clusters* e *outliers* das aglomerações localizadas na zona norte da cidade.**







### Anexo 33. Relatório do processo de agregação de dados através da ferramenta *Space time cube*.

Create Space Time Cube

(Space Time Pattern Mining Tools)

Completed Today at 9:30:08 PM

Parameters

Input Features	E:\tesis\Mapas\caop_15\base_15\sample_fotos_cb2.shp
Output Space Time Cube	C:\Users\ceg\Desktop\emerging\cubo_1.nc
Time Field	upload_dat
Template Cube	
Time Step Interval	1 Months
Time Step Alignment	END_TIME
Reference Time	
Distance Interval	50 Meters
Summary Fields	
Aggregation Shape Type	HEXAGON_GRID

Messages

```

----- Space Time Cube Characteristics -----
Input feature time extent      2007-01-01 00:00:00
                               to 2014-07-31 00:00:00

Number of time steps           91
Time step interval             1 month
Time step alignment            End

First time step temporal bias   3.23%
First time step interval       after
                               2006-12-31 00:00:00
                               to on or before
                               2007-01-31 00:00:00

Last time step temporal bias    0.00%
Last time step interval       after
                               2014-06-30 00:00:00
                               to on or before
                               2014-07-31 00:00:00

Cube extent across space      (coordinates in meters)
Min X                          -95327.3668
Min Y                          -107941.9529
Max X                          -83145.2761
Max Y                          -96316.9529
Rows                           232
Columns                         281
Total bins                     5932472

----- Overall Data Trend - COUNT -----
Trend direction                Decreasing
Trend statistic                 -3.7509
Trend p-value                  0.0002
Completed script CreatespaceTimeCube...

```